CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

DESENVOLVIMENTO DO DICIONÁRIO LIWC 2015 EM PORTUGUÊS DO BRASIL

Flavio Matias Damasceno de Carvalho

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador: Gustavo Paiva Guedes e Silva, D.Sc.

Rio de Janeiro,

Março 2019

DESENVOLVIMENTO DO DICIONÁRIO LIWC 2015 EM PORTUGUÊS DO BRASIL

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Flavio Matias Damasceno de Carvalho

Banca Examinadora:

Presidente, Professor D.Sc. Gustavo Paiva Guedes e Silva (CEFET/RJ) (Orientador)

Professor D.Sc. Joel André Ferreira dos Santos (CEFET/RJ)

Professor Ph.D. Eduardo Soares Ogasawara (CEFET/RJ)

Professor D.Sc. Lilian Vieira Ferrari (UFRJ)

Rio de Janeiro,

Março 2019

CEFET/RJ – Sistema de Bibliotecas / Biblioteca Central

C331 Carvalho, Flavio Matias Damasceno de
Desenvolvimento do dicionário LIWC 2015 em português do Brasil / Flavio Matias Damasceno de Carvalho.— 2019.
69f. + apêndice : il. (algumas color). , grafs. ; enc.

Dissertação. Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2019.
Bibliografia : f. 60-69
Orientador : Gustavo Paiva Guedes e Silva
Inclui apêndice

1. Processamento de linguagem natural (Computação). 2. Mineração de dados (Computação). 3. Algorítmos computacionais. 4. Linguistic Inquiry and Word Count. I. Silva, Gustavo Paiva Guedes e (Orient.). II. Título.

CDD 006.312

Elaborada pelo bibliotecário Leandro Mota de Menezes CRB-7/5281

DEDICATÓRIA

Esse trabalho é dedicado aos meus filhos Felipe, Leandro, Fernando e Leonardo, que foram inspiração para persistir e superar os desafios.

AGRADECIMENTOS

O presente trabalho foi desenvolvido com o apoio da Coordenação de Aperfeicionamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Mesmo que não consiga fazer justiça a todas as pessoas que merecem agradecimento nominalmente nesta página, posso tentar pelo menos deixar registrado minha gratidão àquelas que estiveram mais próximas, tanto referente ao espaço quanto ao tempo de desenvolvimento deste trabalho. Deixo, assim, meus agradecimentos a todos que direta ou indiretamente fizeram parte da minha formação e contribuíram para a pessoa que sou hoje.

Agradeço ao Prof. D.Sc. Gustavo Paiva Guedes e Silva pelo empenho e sentido prático com que sempre me orientou neste trabalho e também pela paciência, conselhos, pela amizade e momentos únicos até aqui.

Agradeço ao Prof. D.Sc. Eduardo Ogasawara, pela coordenação do PPCIC, por todo o primor e esmero que reflete na qualidade deste Programa com perfil pioneiro no Brasil.

Ao corpo docente do CEFET-RJ, agradeço pelos ensinamentos que passaram durante o mestrado, e especialmente para Prof.^{a} D.Sc. Kele Belloze e Prof. D.Sc. Joel dos Santos pela organização dos seminários, sem dúvida muito importantes para fortalecer os conteúdos vistos em sala de aula e adquirir novos conhecimentos, além de propiciar a oportunidade de ficar a par das atividades em termos de ensino, pesquisa e extensão relacionados à Computação.

Agradeço à Prof.^{a} D.Sc. Lilian Ferrari pelo apoio e revisões do dicionário.

Agradeço igualmente a todos os meus colegas do Mestrado em Ciência da Computação pelo apoio, amizade e também pela boa vontade de ajudar em vários momentos. Especialmente a M.Sc. Rafael Rodrigues, Yukio Okuno, Gabriel Santos, Carlos Teles, Francimary Oliveira, Adalberto Andrade, Raphael Martins, tanto pela oportunidade de juntos publicarmos artigos, quanto pelos trabalhos de avaliação desenvolvidos em conjunto.

Agradeço à minha família por serem tão importantes na minha vida, especialmente a minha esposa Léa pela paciência, compreensão, alegria e amor. Também agradeço aos meus sogros Jair e Selma pela constante presença, companheirismo e apoio que me deram.

RESUMO

Desenvolvimento do Dicionário LIWC 2015 em Português do Brasil

Atualmente, há muitos textos em formato digital, sendo que técnicas e metodologias da área de Mineração de Texto podem ser usadas para obter informações úteis desses textos. Uma dessas metodologias analisa textos com o Linguistic Inquiry and Word Count (LIWC), um programa que possui várias versões que foram melhoradas ao longo dos anos. Este programa pode utilizar tanto um arquivo de dicionário padrão, quanto versões do dicionário padrão em outros idiomas. Avaliações mostram que utilizar o dicionário em português, que foi baseado no dicionário em inglês da versão 2007, traz resultados pouco satisfatórios para detecção de valência negativa em textos. Também são encontrados erros ortográficos e problemas relacionados à categorização neste dicionário, o que impacta negativamente resultados em análise de textos. Reconhecendo a necessidade de métodos para analisar texto na língua portuguesa e, não tendo conhecimento do desenvolvimento de uma versão mais recente em português, iniciamos o desenvolvimento de uma nova versão em português do dicionário para o LIWC. Trabalhamos com o conjunto de palavras disponíveis na versão de 2015 em inglês e produzimos um novo dicionário compatível com a última versão disponível do programa. Para verificar o desempenho em tarefas de classificação, realizamos experimentos para classificar: (i) autores de textos e (ii) conteúdo das publicações nas redes sociais de acordo com a polaridade do sentimento. As medidas utilizadas para avaliar os resultados obtidos pelos algoritmos de classificação empregados apresentaram valores maiores na nova versão em português do dicionário, comparando com o dicionário de 2007. Esses experimentos sugerem que ajustar palavras de forma adequada a categorias linguísticas e psicológicas melhora resultados nas tarefas associadas às áreas de Computação Afetiva e Análise de Sentimentos.

Palavras-chave: Dicionário; Mineração de Textos; Computação Afetiva; Análise de Sentimentos

ABSTRACT

Development of the Portuguese LIWC 2015 Dictionary

Currently, there are many texts in digital format, and techniques and methodologies from the Text Mining field can be used to obtain useful information of these texts. One of these methodologies analyzes texts with the Linguistic Inquiry and Word Count (LIWC), a program that has several versions that have been improved over the years. This program can use both a standard dictionary file and also versions of the standard dictionary in other languages. Evaluations show that using the Portuguese dictionary, which was based on the 2007 English version of the dictionary, brings unsatisfactory results for detection of negative valence in texts. Spelling errors and problems related to categorization in this dictionary are also found, which negatively impacts results in text analysis. Acknowledging the need for methods to analyze text in the Portuguese language and, not aware of the development of a more recent version in Portuguese, we started the development of a new Portuguese version of the LIWC dictionary. We worked with the set of words available in the 2015 version in English and produced a new dictionary compatible with the latest available version of the program. To verify performance in classification tasks, we performed experiments to classify: (i) text authors and (ii) the content of publications in social networks according to sentiment polarity. The measures used to evaluate the results obtained by the classification algorithms employed presented higher values in the new Portuguese version of the dictionary, comparing with the 2007 dictionary. These experiments suggest that adjusting words appropriately to linguistic and psychological categories improves results in tasks associated with the areas of Affective Computing and Sentiment Analysis.

Keywords: Lexicon; Text Mining; Affective Computing; Sentiment Analysis

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AGS Aspectos Gramaticais Ou Semânticos

A-P D'Agostino-Pearson

API Interface De Programação De Aplicação

AS Análise De Sentimentos

BT Bing Tradutor

CA Computação Afetiva

DT *Decision Tree*

GT Google Tradutor

LIWC *Linguistic Inquiry and Word Count*

LMT *Logistic Model Tree*

MQD Meu Querido Diário

MT Mineração De Textos

NB *Naive* Bayes

NBM *Naive* Bayes Multinomial

RF *Random Forest*

S-W Shapiro-Wilk

VOP Vocabulário Ortográfico Do Português

SUMÁRIO

Introdução

Uma grande variedade e quantidade de textos são escritos e armazenados em formato digital devido ao desenvolvimento e disseminação de dispositivos computacionais. Nos mais diversos ambientes se observa o uso de dispositivos computacionais como ferramenta básica e essencial para a escrita de documentos, relatórios técnicos, roteiros de teatro, planos de aula, livros, entre outros. Ao mesmo tempo as obras de séculos passados, anteriormente disponíveis apenas em livros de papel, foram digitalizadas para se aproveitar o formato digital (BURK, 2008).

Podemos ainda apontar que é possível registrar opinião pessoal sobre notícias, produtos e serviços em formulários disponíveis em *websites*. Além disso, pessoas com acesso a redes sociais e aplicativos de mensagens têm registrado muitos eventos, sentimentos e emoções em formato de texto. Dessa forma, é notável a grande quantidade de texto em formato digital que pode ser compartilhado, enviado e processado por ferramentas computacionais.

As pessoas utilizam a linguagem e podem expressar de formas diferentes um mesmo conteúdo ou tema, sendo que o processamento por ferramentas computacionais torna viável evidenciar diversos destes aspectos em textos. Diante deste cenário, vemos a Mineração de Textos (MT) apresentar metodologias computadorizadas de tratamento dos dados em formato de texto, revelando informações interessantes tanto sobre quem escreve (H. LIU; KEŠELJ, 2007), quanto sobre o conteúdo, como no caso de análises de documentos de patentes (TSENG; C.-J. LIN; Y.-I. LIN, 2007). Por isso, as operações de MT utilizando ferramentas de análise de textos podem trazer informações relevantes a diversas áreas de conhecimento (JELIER et al., 2008).

Uma das ferramentas informatizadas disponíveis para análise de texto é o programa *Linguistic Inquiry and Word Count* (LIWC), que se baseia em um dicionário[1] composto por palavras organizadas em categorias (PENNEBAKER; FRANCIS; BOOTH, 2001). O LIWC foi desenvolvido com o objetivo de analisar os componentes emocionais, cognitivos e estruturais de textos de acordo com essas categorias, considerando a quanti-

[1] Neste trabalho, utilizamos os termos 'léxico' e 'dicionário' para nos referirmos a conjuntos de termos próprios ou de vocábulos de uma língua.

dade de palavras que o programa encontra nos textos. A versão mais recente do arquivo de dicionário do LIWC em inglês foi lançada junto com uma nova versão do programa em 2015 (PENNEBAKER et al., 2015a). Nesse estudo, essa versão do dicionário em inglês é denominada LIWC_2015en.

Existem versões anteriores do dicionário do LIWC em inglês para diferentes idiomas. Versões baseadas na versão de 2007 do dicionário do LIWC em inglês contêm palavras em 64 categorias organizadas hierarquicamente. Quanto ao LIWC_2015en, que conta com 73 categorias, existem versões apenas para Alemão, Chinês e Holandês, até onde temos conhecimento.

Para a língua portuguesa, que é uma das línguas mais faladas no mundo (SIMONS; FENNIG, 2017) e uma das dez línguas mais utilizadas na internet[2], existe uma versão do dicionário para o português do Brasil, baseada na versão de 2007 do LIWC (FILHO; PARDO; ALUÍSIO, 2013). Nesse trabalho, essa versão é denominada LIWC_2007pt.

O LIWC tem sido utilizado em estudos que analisam textos em diversas línguas para obter informações sobre gênero (SCHLER et al., 2006; GUEDES et al., 2016), detectar doenças mentais (SHIBATA et al., 2016) e ajudar no diagnóstico de distúrbios afetivos como depressão (RUDE; GORTNER; PENNEBAKER, 2004). Também encontramos estudos utilizam técnicas de MT e o LIWC para verificar aspectos da sociedade (CAETANO et al., 2017; PETTIJOHN; SACCO JR, 2009), marketing (PETTIJOHN; SACCO JR, 2009; NASSIRTOUSSI et al., 2014), saúde (SHIBATA et al., 2016; RUDE; GORTNER; PENNEBAKER, 2004) e vida cotidiana (SCHLER et al., 2006; CARVALHO et al., 2018), mostrando desde resultados úteis para prever eleições (TUMASJAN et al., 2010) quanto tendências de mercado (NASSIRTOUSSI et al., 2014; H. LIU; KEŠELJ, 2007). Utilizando textos processados com o LIWC junto com algoritmos de classificação, também é possível inferir a idade dos usuários nas redes sociais (SCHLER et al., 2006; CARVALHO et al., 2018), o que pode ter como aplicação final a identificação de possíveis predadores sexuais (RODRIGUES et al., 2017a). Além disso, é possível obter informações sobre inclinações políticas (CAETANO et al., 2017) e status social e econômico (PETTIJOHN; SACCO JR, 2009).

Considerando a grande variedade de possíveis aplicações e estudos sendo realizados, o número de citações ao LIWC_2007pt também vem aumentando nos últimos anos, o que pode ser verificado com a utilização do Google Acadêmico[3]. Uma pesquisa

[2] http://www.internetworldstats.com/stats7.htm

[3] `https://scholar.google.com.br/`

efetuada no momento em que este texto é escrito retorna 44 citações exclusivas. Na Figura 1, observamos a contagem de citações por ano entre 2013 e 2018, sendo importante mencionar que em 2018, 16 obras citam o LIWC_2007pt.

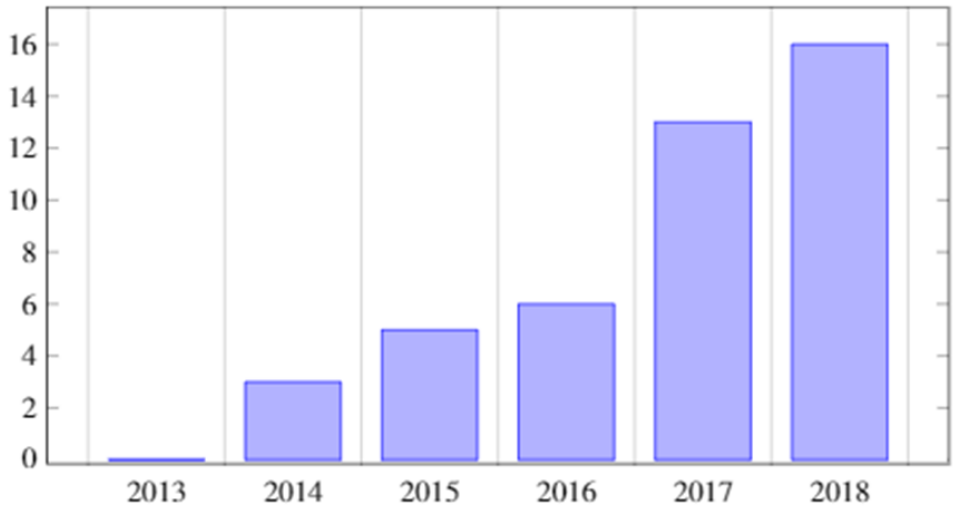

Figura 1 – Resultado de consulta ao Google Acadêmico sobre a quantidade anual de trabalhos citando o LIWC_2007pt.

Entendemos que este aumento no número de citações sugere uma importância crescente deste recurso em estudos acadêmicos em português, no entanto, também começam a ser encontrados diversos problemas com esse dicionário. Nas primeiras avaliações publicadas com o LIWC_2007pt foi observado bom desempenho alcançado na detecção de valência positiva em textos na língua portuguesa (FILHO; PARDO; ALUÍSIO, 2013). Porém, o desempenho observado na detecção de valência negativa não foi tão bom (FILHO; PARDO; ALUÍSIO, 2013). Na documentação disponível sobre o processo de desenvolvimento do LIWC_2007pt, afirma-se que nenhuma revisão do trabalho manual da tradução foi feita, e também que esta pode ser melhorada[4].

Acreditamos que esse impacto negativo nos resultados possa estar relacionado a problemas que encontramos durante o desenvolvimento de diversos estudos. É possível identificar no LIWC_2007pt que algumas palavras não estão nas categorias apropriadas, enquanto outras estão incorretamente associadas a algumas categorias (CARVALHO et al., 2018). Outra questão relevante está relacionada às inconsistências devido a erros de ortografia e palavras usando o caractere curinga '*' (CARVALHO; SANTOS; GUEDES, 2018). Detalhamos na Seção 2.2 as questões relacionadas aos problemas do LIWC_2007pt e que reforçaram que existe uma lacuna que pode ser preenchida com a

[4] http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc, conforme acessado em 3 de maio de 2019.

criação de um novo dicionário em português do Brasil.

Somando-se a questões desta natureza, que podem ser tratadas, há também o fato de que existe, conforme mencionado, uma versão mais recente do LIWC para a língua inglesa. O LIWC_2015en foi desenvolvido após a tradução para português (i.e., LIWC_2007pt) ter sido disponibilizada em 2011. O LIWC_2015en introduz várias novas categorias, buscando a melhoria e refinamento dos resultados do programa. Para utilizar os recursos da versão de 2015 do programa LIWC para análise de textos em português, uma versão em português do dicionário com a mesma estrutura e categorias do LIWC_2015en precisa estar disponível, mas não temos conhecimento sobre o desenvolvimento de uma até agora.

Nesse cenário, reconhecemos que é necessário o desenvolvimento de métodos para análise de textos em outras línguas, além do inglês, especialmente em português (REIS et al., 2015; RAMÍREZ-ESPARZA et al., 2007). Por essas razões, este trabalho apresenta os esforços para desenvolver uma versão de um arquivo de dicionário em português do Brasil com a mesma estrutura e categorias do LIWC_2015en. Nesse estudo, criamos um novo dicionário, denominado LIWC_2015pt, em etapas. Essas etapas consistem na criação de dicionários de menor tamanho (i.e., com subconjunto de categorias) e, em um passo seguinte, há a consolidação desses dicionários em um único arquivo.

Observamos que a tarefa de criar novos dicionários por meio de modificações baseadas no LIWC_2007pt se mostra viável apenas para categorias que são bem definidas ou com número pequeno de palavras, como algumas categorias referentes a classes gramaticais. Considerando os recursos disponíveis, avaliamos ser melhor desenvolver novos dicionários baseados no LIWC_2015en para as categorias mais amplas (e.g., adjetivos, verbos), por esse dicionário já ter sido submetido a processos de revisão e correção ao longo das atualizações. Analisar o LIWC_2007pt para encontrar todas as palavras com problemas ortográficos, corrigir a categorização de todas as palavras e então buscar por novas palavras adequadas a cada categoria consistiria em uma tarefa bem mais custosa. Uma forma de exemplificar esta questão é observando as categorias que representam respostas emocionais ou afetivas no LIWC_2007pt, que contêm um número de palavras que excede 20 vezes o número de palavras nas categorias afetivas do LIWC_2015en.

Considerando a importância dos pronomes para Computação Afetiva (CA) e

Análise de Sentimentos (AS) (PENNEBAKER, 2011a; OFEK et al., 2015), iniciamos o desenvolvimento do LIWC_2015pt com a criação de um novo dicionário de pronomes, denominado PronounBP (CARVALHO et al., 2018). Usamos como base a categoria dos pronomes do LIWC_2007pt, que contém 128 palavras, incluindo suas subcategorias. Criamos também um dicionário com palavras que representam respostas emocionais ou afetivas, denominado AffectPT-br, e incluímos as subcategorias de emoções positivas e negativas do LIWC_2015en (CARVALHO; SANTOS; GUEDES, 2018).

Desenvolvemos então outros dicionários com demais categorias do LIWC_2015en, fazendo uso tanto do processo de criação desse dicionário afetivo em português do Brasil, i.e. baseado na tradução das palavras conforme organizadas no LIWC_2015en, como também utilizando listas de palavras adequadas às categorias, como no processo de criação do dicionário de pronomes. Na etapa de integração, expandimos os dicionários no sentido em que diminuímos o uso das palavras terminadas com um caractere curinga[5], de forma a incluir variações de gênero e número e assim obter consistência com diferentes palavras nas categorias de aspectos gramaticais, como adjetivos e advérbios. Também incluímos os verbos conjugados, permitindo a diferenciação de formas com foco no presente, passado e no futuro.

A versão em português do Brasil do dicionário que apresentamos neste trabalho possui um total de 14.459 palavras em 73 categorias. É maior que a versão em inglês (6.400), o que é esperado pela inclusão de palavras com variações de gênero, número, grau, etc. Vale ressaltar que um número muito maior de palavras do LIWC_2007pt (127.000) pode não representar um impacto positivo no número de palavras a serem identificadas e contadas nos textos (CARVALHO; SANTOS; GUEDES, 2018), considerando os problemas já mencionados e que serão melhor detalhados na Seção 2.2.

Considerando estes aspectos, este trabalho tem objetivo de melhorar a classificação de textos em português em tarefas de MT utilizando o LIWC, pelo desenvolvimento de um dicionário em português para uso na versão 2015 do LIWC. Os objetivos específicos incluem (i) a coleta de amostras públicas de textos, como *corpus* paralelo e bidirecional[6] de português e inglês e publicações de redes sociais, (ii) a análise computadorizada desses textos e (iii) a avaliação empírica dos resultados. Executamos a avaliação usando

[5] O LIWC usa o '*' (asterisco) para indicar a aceitação de todas as letras, hifens ou números após a sua aparição em uma palavra.

[6] Um tipo de conjunto de dados em duas línguas, com textos originais e suas respectivas traduções conectadas frase a frase (BAKER, 1995).

análise estatística para comparar a versão referência em português, o LIWC_2007pt, com o dicionário desenvolvido, e aplicando ambos em tarefas de classificação.

Desta maneira, essa dissertação possui como principais contribuições:

- O desenvolvimento de um dicionário em português com todas as categorias de palavras encontradas no LIWC_2015en;
- A avaliação empírica de resultados de comparações entre as versões existente e a que foi desenvolvida, consistindo em:

  - A realização de uma comparação estatística dos resultados de análises de diferentes versões de dicionários do LIWC: LIWC_2015en e LIWC_2015pt

  - A avaliação de desempenho pelo uso de textos processados com o LIWC_2007pt e o LIWC_2015pt em tarefas de classificação utilizando MT.

Ao longo do desenvolvimento deste trabalho, se encontram registradas contribuições com publicações de estudos utilizando o LIWC:

- Carvalho, Flavio, and Guedes, Gustavo Paiva. *Night Sleep Deprivation: Computational Analysis of Language Effects.* In: Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web (WebMedia 2017), Gramado.
- Okuno, H.Y., and Carvalho, F., and Guedes, G.P., and Torres, M., *ATAnalysis-Toward a psycholinguistic method to analyze video textual information.* In: Latin America Data Science Workshop - (LADaS 2018), Rio de Janeiro.
- Rodrigues, Rafael Guimarães, and Carvalho, Flavio, and Guedes, Gustavo Paiva. *Towards the creation of a tool for identifying personality in virtual learning environments.* In: Latin American Conference on Learning Technologies (LACLO 2018), São Paulo.
- Carvalho, Flavio and Rodrigues, Rafael Guimarães, and Ferrari, Lilian, and Guedes, Gustavo Paiva. *LIWBC: a bigram algorithm to enhance results in polarity classification.* In: Proceedings of the 24th Brazillian Symposium on Multimedia and the Web (WebMedia 2018), Salvador.
- Carvalho, Flavio and Santos, Gabriel, and Guedes, Gustavo Paiva. *AffectPT-br: an Affective Lexicon based on LIWC 2015*. In: 37th International Conference of the Chilean Computer Science Society (SCCC 2018), Santiago – Chile.

- Carvalho, Flavio and Rodrigues, Rafael Guimarães, and Ferrari, Lilian, and Guedes, Gustavo Paiva. *A dictionary of pronouns for Brazilian Portuguese*. In: Congresso Internacional de Informática Educativa (TISE 2018), Brasília.

Além deste capítulo introdutório, este trabalho está estruturado em mais seis capítulos que tratam da fundamentação teórica, trabalhos relacionados, metodologia, resultados, considerações e cenários futuros. O capítulo 1 aborda o referencial necessário para o entendimento dos assuntos relacionados ao desenvolvimento deste trabalho. Desta forma, trata de informações básicas como conceitos, motivações, contribuições e áreas relacionadas à análise de texto informatizado.

No capítulo 2 apresentamos trabalhos relacionados ao estudo desta dissertação, começando com o LIWC. Destacamos nesta seção as informações encontradas sobre o LIWC_2007pt. Trazemos ainda uma relação de trabalhos que abordam a tradução de dicionários do LIWC para outros idiomas.

O capítulo 3 trata diretamente das contribuições deste trabalho. Inicialmente descrevemos como desenvolvemos o LIWC_2015pt em etapas, detalhando a criação de dicionários com categorias selecionadas do LIWC_2015en. Em seguida, nesse capítulo, listamos a criação e integração dos arquivos de dicionário em um único arquivo.

Os resultados são apresentados no capítulo 4, no qual trazemos comparação de resultados em tarefas de classificação, pela análise de entradas de texto de diferentes redes sociais. Mostramos também uma comparação dos resultados obtidos de análises de um *corpus* paralelo. Ainda apresentamos uma comparação do tempo de processamento dos conjuntos de textos utilizando o LIWC_2007pt e o LIWC_2015pt.

Nas considerações finais trazemos uma discussão sobre os resultados obtidos com o uso do LIWC_2015pt. Também abordamos algumas dificuldades encontradas e tecemos considerações acerca das avaliações experimentais. Finalmente, apresentamos cenários futuros e possíveis contribuições, abordando a expectativa para continuidade dos estudos.

1- Fundamentação Teórica

Neste capítulo, apresentamos os conceitos das áreas que estão relacionados com os temas no trabalho. Tanto a CA quanto a AS podem utilizar técnicas de mineração de textos e de aprendizagem de máquina para analisar o sentimento do autor ou do que está sendo escrito. Por isso, essa fundamentação serve para ajudar a compreensão do conteúdo dos capítulos seguintes.

Na Seção 1.1, abordamos técnicas utilizadas e exemplos de uso da MT. Em seguida abordamos, na Seção 1.2, estudos que exemplificam a aplicação do conhecimento dos campos da CA e AS. Na Seção 1.3, discorremos sobre o estudo de sentimentos e emoções em textos usando dicionários afetivos e, na Seção 1.4, trazemos de forma introdutória cada algoritmo de aprendizagem de máquina que utilizamos para classificação.

1.1- Mineração de Textos

A MT é uma área influenciada pela área de mineração de dados (FELDMAN; SANGER, 2007). De maneira mais abrangente, a mineração de dados usa análise matemática para extrair padrões e tendências, empregando recursos computacionais para obter informações úteis sobre grandes quantidades de dados estruturados (HAN; PEI; KAMBER, 2011). A MT se concentra na busca de padrões válidos, novos, potencialmente úteis e compreensíveis a partir de dados não estruturados, que se apresentam como documentos textuais (FAYYAD et al., 1996a).

Na MT se obtém as informações dos textos pela identificação de padrões e tendências usando algoritmos e métodos da área de aprendizado de máquina e estatística (SATHYA; RAJENDRAN, 2015). As palavras, pontuações e estilo do texto trazem consigo diversos dados dos sentimentos de quem o está escrevendo e não apenas a mensagem principal que se quer passar pelo texto (PENNEBAKER, 2011b). O objetivo é concentrado na representação de textos em espaços vetoriais, desenvolvimento de algoritmos de classificação e desenvolvimento de aplicações direcionadas para AS (SATHYA;

RAJENDRAN, 2015; MIKOLOV et al., 2013).

O processo de MT pode ser dividido nas fases de pré-processamento, núcleo de mineração de texto e apresentação (FAYYAD; PIATETSKY-SHAPIRO; SMYTH et al., 1996b). No pré-processamento, é realizada uma análise linguística, sendo os documentos processados para se encontrar padrões e relacionamentos (FELDMAN; SANGER, 2007). Ou seja, se trata de converter os textos para um formato mais adequado à manipulação computacional.

Na etapa 'núcleo de mineração de texto', o objetivo é extrair padrões que devem ser capazes de trazer informações novas, úteis e compreensíveis sobre os dados (FELDMAN; SANGER, 2007). Com base no que foi obtido na análise linguística, avaliações estatísticas ou técnicas que usam aprendizado de máquina são utilizadas para descoberta de padrões, análise de tendências e de descoberta de conhecimento. Também ocorrem, nesta etapa, comparações entre alguns desses padrões encontrados.

Na fase de apresentação, as descobertas do processamento computacional são traduzidas para a linguagem humana natural (FELDMAN; SANGER, 2007). A pesquisa de atributos nos documentos e o uso de recursos de imagens, como diagramas de árvores, são feitas nessa fase para evitar a verificação de uma enorme quantidade de documentos de forma manual. Dessa forma se consegue ter uma boa ideia do que pode ser encontrado em dada coleção.

Estudos utilizando metodologias apoiadas em MT têm aplicações nos eixos governamental, científico e empresarial. De acordo com a análise realizada ou com a função de negócios, podem ser feitas diferentes classificações em categorias que empregam técnicas desta área. Dentre estas categorias, temos Inteligência Competitiva (*Enterprise Business Intelligence*) (CHITICARIU; LI; REISS, 2013), gerenciamento de registros (*E-Discovery*) (GROSSMAN; CORMACK, 2010), Segurança Nacional ou Inteligência (SAMANTA; PANCHAL, 2016), Descoberta científica (COHEN; HUNTER, 2008), Ferramentas de AS (BRADLEY; LANG, 1999; PENNEBAKER et al., 2015a), Serviços ou ferramentas de linguagem natural (MANNING; SCHÜTZE, 1999), entre outras.

1.2- Computação Afetiva e Análise de Sentimentos

A emoção desempenha papel fundamental na percepção, memória e atenção, na tomada racional de decisões e na interação homem-máquina (PICARD et al., 1995). A CA e AS são campos científicos envolvidos com tarefas computacionais para tentar interpretar as emoções humanas. A diferença entre o uso dos termos é que, enquanto normalmente se usa 'Análise de Sentimentos' para se identificar apenas a polaridade (e.g., na frase de um comentário), se usa 'Computação Afetiva' para a identificação das diferentes emoções (CAMBRIA et al., 2013).

É destacado como um dos precursores da CA o trabalho de Manfred Clynes (PICARD, 2010), que construiu um dispositivo capaz de medir emoções como tristeza, raiva, alegria, entre outras. Na literatura acadêmica, (KLEINE-COSACK, 2008) reconhece que esse campo de conhecimento se desenvolveu a partir da publicação em 1995 do livro 'Affective Computing' por Picard e colaboradores. A aplicação dos conhecimentos dessa área permitem a identificação de opiniões e a classificação da orientação ou polaridade destas e, portanto, a apresentação dos resultados de forma agregada e resumida (BECKER; TUMITAN, 2013).

Nos dias de hoje, existe um grande e diverso conjunto de redes sociais e aplicativos de troca de mensagens disponíveis e amplamente difundido entre os usuários (ALTHOFF; JINDAL; LESKOVEC, 2017). Considerando também a disseminação e uso de dispositivos portáteis de comunicação, constantemente ao alcance de seus usuários (BERENGUER et al., 2017), se observa a possibilidade de acesso à internet a todo momento para uso de tais aplicativos. Com isso, temos um cenário em que se mostra possível que ocorram registros escritos em várias ocasiões ao longo de um dia.

Estes registros escritos podem servir para estudo das emoções, sendo que, pelo volume produzido, é uma importante modalidade de detecção de estados afetivos dos usuários (VALITUTTI; STRAPPARAVA; STOCK, 2004). As medidas realizadas envolvendo variáveis que representam emoções trazem informações que são de interesse em diversas áreas, tais como ciências de gestão, ciência política, economia e ciências sociais (B. LIU, 2012). Um estudo de análise textual pode trazer informações sobre o status social e econômico (PETTIJOHN; SACCO JR, 2009), diferenças culturais (NAKAYAMA; WAN, 2018), idade (RODRIGUES et al., 2017a), gênero (GUEDES et al., 2016), entre outras ca-

racterísticas. Também é útil na detecção de doenças mentais e para ajudar o diagnóstico (RUDE; GORTNER; PENNEBAKER, 2004; SHIBATA et al., 2016; DE CHOUDHURY; COUNTS; HORVITZ, 2013), e também em campanhas de marketing e atividades forenses (BOYD; PENNEBAKER, 2015). Estes exemplos ilustram a possibilidade de aplicações nos eixos governamental, científico e empresarial.

Usando conhecimentos da área de AS, podem ser obtidos dados relacionados com o sentimento de usuários sobre um determinado assunto. Uma das formas mais comuns de realizar essa análise é pela da polaridade de sentimentos (AGARWAL; BIADSY; MCKEOWN, 2009). Dessa maneira, por exemplo, palavras associadas a elogios, comparações positivas e relatos de bons acontecimentos são 'positivas'. Críticas, relatos de experiências ruins, comparações negativas são 'negativas'. Alguns métodos tratam a polaridade como um resultado discreto binário (positivo ou negativo) (PRABOWO; THELWALL, 2009) ou ternário (positivo, negativo ou neutro) (AGARWAL; BIADSY; MCKEOWN, 2009).

A expressão emocional em redes sociais permite também o monitoramento da saúde mental populacional observando publicações de textos relacionadas à depressão (MOWERY; BRYAN; CONWAY, 2017). A análise computacional de sentimentos e emoções expressos em textos nestas redes também possibilita realizar predições sobre satisfação no trabalho, sucesso profissional e relacionamento romântico, o que vêm motivando estudos de inferência de fatores da personalidade (GOLBECK; ROBLES; TURNER, 2011). Estes casos são úteis para exemplificar o estudo de sentimentos e emoções em textos usando um recurso conhecido como dicionários afetivos.

1.3- Dicionários Afetivos

Para detecção e classificação automática de emoções, uma das abordagens consiste em realizar a análise computacional de sentimentos e emoções em textos utilizando aprendizado de máquina (SEBASTIANI, 2002; IKONOMAKIS; KOTSIANTIS; TAMPAKAS, 2005). Porém, a utilização destas técnicas encontra dificuldades como na aplicabilidade de modelos que geralmente são bem restritos ao contexto para o qual foram criados, pela necessidade de boa quantidade de dados previamente validados para

treinamento (PANG; L. LEE; VAITHYANATHAN, 2002), além da atenção que deve ser dada ao escolher os dados de treinamento de forma a se evitar o viés no classificador. Uma alternativa é utilizar métodos léxicos, que utilizam Dicionários Afetivos, i.e. listas e dicionários de palavras associadas a sentimentos (TABOADA et al., 2011), e desta forma apresentam a vantagem de não dependerem de dados rotulados para treinamento (BENEVENUTO; RIBEIRO; ARAÚJO, 2015).

Ao realizar uma análise utilizando um dos métodos léxicos, e.g. utilizando o LIWC, primeiramente é feita a seleção de pelo menos um arquivo de texto para processamento pelo programa, que cria uma representação vetorial do texto analisado. Nesse vetor de saída, cada posição é incrementada pela identificação de atributos do texto como quantidade total de palavras no texto encontradas no dicionário, elementos de pontuação, palavras por frase, porcentagem de palavras do texto com mais de seis letras, além das palavras em cada categoria. Para cada arquivo de texto, a versão de 2015 do LIWC, usando o dicionário padrão (LIWC_2015en), grava aproximadamente 90 variáveis organizadas em colunas em um arquivo de saída.

Neste arquivo de saída, o LIWC registra os dados do nome do arquivo, contagem de palavras, 4 variáveis de linguagem sumarizadas (pensamento analítico, influência, autenticidade e tom emocional), e o resultado da contagem para as diversas categorias. Dessas, 3 são categorias descritoras gerais (palavras por frase, porcentagem de palavras-alvo capturadas pelo dicionário e porcentagem de palavras no texto com mais de seis letras), 21 dimensões linguísticas padrão (e.g., porcentagem de pronomes, artigos, etc.) e também 41 categorias de palavras de construtos psicológicos (e.g., afeto, cognição, impulsos). Além dessas, são registrados ainda valores de 6 categorias de interesse pessoal (e.g., trabalho, casa, lazer), 5 categorias de linguagem informal (aprovação, preenchimentos, palavrões, *internetês*) e 12 categorias de pontuação (pontos, vírgulas, etc) (PENNEBAKER et al., 2015a).

Este arquivo de saída com a representação vetorial do texto pode então ser utilizada como dados de entrada para algoritmos de aprendizado de máquina. Isto porque, desta maneira, uma determinada seleção de documentos de texto está representada em um formato estruturado. Em uma tarefa de classificação, algoritmos usam modelos identificando a representatividade de cada atributo para cada classe, e.g. categoria do dicionário, em relação às posições do vetor do texto analisado, de acordo com métricas indicadas na Seção 1.4.

1.4- Algoritmos de classificação

Em aplicações de MT, uma tarefa de classificação busca identificar a qual classe cada item de uma amostra de texto pertence. Nos experimentos que detalhamos na Seção 4.1, a comparação do desempenho em tarefas de classificação é feito observando medidas usando diferentes algoritmos. Para isso, os algoritmos funcionam de forma a aprender a importância de categorias, observando os valores em cada uma nas classes. Baseado na literatura, conforme indicado na Seção 4.1, escolhemos os algoritmos *Naive* Bayes (NB), *Naive* Bayes Multinomial (NBM), *Decision Tree* (DT), *Random Forest* (RF), *Logistic Model Tree* (LMT) e J48.

O algoritmo NB se baseia na teoria de decisão Bayesiana (LANGLEY; IBA; THOMPSON et al., 1992), que indica a probabilidade de um evento dado o conhecimento prévio relacionado ao evento. O cálculo é realizado conforme a Equação 1, onde **C** indica as classes e A os atributos (no nosso caso, as categorias de palavras). O NBM utiliza uma variação dessa teoria, e usa um modelo multinomial para obter a frequência de uma variável no conjunto (MCCALLUM; NIGAM et al., 1998).

$$P(A|\mathbf{C}) = P(A)\frac{P(\mathbf{C}|A)}{P(\mathbf{C})} \tag{1}$$

Os algoritmos DT, RF, LMT e J48 utilizam o método conhecido como 'Árvore de decisão' ou simplesmente DT (QUINLAN, 1986), que classifica instâncias usando uma estrutura de ordenação da 'raiz' para algum nó 'folha', onde cada nó da árvore representa um atributo. O algoritmo RF (HO, 1995) funciona gerando internamente várias DT e combinando o resultado da classificação de todas elas (SVETNIK et al., 2003). O LMT combina aprendizagem usando 'Regressão Logística' (LR) e o método da 'Árvore de decisão' (LANDWEHR; HALL; FRANK, 2005). Já o J48 se trata de uma implementação em Java do algoritmo C4.5 (QUINLAN, 2014), que usa o conceito de entropia da informação para montar as árvores.

Para produzir os experimentos de classificação descritos na Seção 4.1, utilizamos o Weka (HALL et al., 2009) utilizando cada um dos algoritmos com seu conjunto de configurações padrão. Os resultados dos algoritmos de classificação obtidos com o Weka são relatados considerando as medidas de precisão (P) da classificação, revocação (R)

e a medida F, mais conhecida como *F-measure* ou F_1. Desta forma, P representa a porcentagem de itens classificados corretamente, i.e. o número de resultados corretos de uma classe dividido pelo número de todos os resultados atribuídos a essa classe (i.e. que foram classificados corretamente e incorretamente). R representa a sensibilidade do sistema, i.e. a relação do número de resultados corretos de uma classe pelo total de itens que realmente estão nesta classe.

Com os valores obtidos de P e R, obtemos o valor da F_1. Conforme expresso pela fórmula 2, F_1 é uma medida ponderada usando os valores de P e R (VAN RIJSBERGEN, 1979). Os valores de F_1 estão entre 0 e 1, onde 0 indica o pior valor e 1 indica o melhor valor, ou seja, valores perfeitos de P e R.

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{2}$$

Cada experimento com o Weka utiliza a técnica de validação cruzada de particionamentos, denominada validação *k-fold*. Conforme indicado por (KOHAVI et al., 1995), utilizamos *k-fold* com dez partições. Observamos a medida da média do valor de F_1 de cada classe para comparar os resultados obtidos pela classificação utilizando arquivos com a representação vetorial dos textos obtida pelo processamento com cada dicionário.

2- Trabalhos relacionados

Neste capítulo, são descritos trabalhos relacionados ao estudo desta dissertação. Abordamos inicialmente o LIWC, um programa de análise computadorizada de textos que pode ser configurado para utilizar diversos arquivos de dicionários, trazendo informações sobre o seu dicionário padrão. Em seguida, apresentamos as informações que foram possíveis de obter referentes ao desenvolvimento do LIWC_2007pt. Também são mencionados trabalhos que tratam de apresentar o processo de desenvolvimento do LIWC para outros idiomas.

2.1- LIWC

Os conjuntos, tanto de palavras de uma língua quanto de termos próprios de algum campo, são conhecidos como léxicos ou dicionários[1]. Para tarefas de identificação de emoções em textos, os conjuntos de palavras são focados em emoções e chamados de dicionários afetivos. Em sistemas de CA, os dicionários afetivos são parte essencial da realização de tarefas que envolvem o processamento computacional de linguagem (FREITAS, 2013).

O LIWC é um programa de análise computadorizada de texto para medir a quantidade de palavras em um texto, inseridas em pelo menos uma das mais de 60 categorias em que está organizado o seu arquivo de dicionário. O LIWC foi desenvolvido com o objetivo de analisar componentes emocionais, cognitivos e estruturais a partir de textos (PENNEBAKER et al., 2015b). Pode ser dividido em duas partes, uma é o programa principal e a outra é um dicionário, que define quais palavras o programa deve contar nos arquivos de texto a serem analisados.

Versões diferentes do LIWC foram desenvolvidas nos últimos anos, trazendo modificações no programa e no arquivo de dicionário. Desta forma, algumas das catego-

[1] https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/dicionário/, https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/léxico/

rias originais foram removidas e o número de palavras foi alterado. LIWC_2015en, a versão mais recente do arquivo de dicionário, possui um total de 6.400 palavras (PENNEBAKER et al., 2015a).

No dicionário, as palavras estão organizadas em grupos de um domínio específico. Esses agrupamentos são chamados de subdicionários ou, como preferimos utilizar neste trabalho, categorias de palavras (PENNEBAKER et al., 2015b). O dicionário contém palavras designadas a uma ou mais categorias que refletem processos linguísticos, psicológicos ou relacionadas a diversos assuntos, como pronomes (*pronoun*), emoções positivas (*posemo*), processos sociais (*social*) e assim por diante.

No dicionário utilizado pelo LIWC, as principais categorias são divididas em subcategorias, conforme ilustrado na Figura 2. Por exemplo, a categoria de pronomes é dividida em duas subcategorias: pronomes pessoais (*ppron*) e pronomes impessoais (*ipron*). A subcategoria *ppron* é dividida em 5 mais subcategorias: primeira pessoa do singular (*i*), primeira pessoa do plural (*we*), segunda pessoa (*you*), terceira pessoa do singular (*shehe*) e terceira pessoa do plural (*they*).

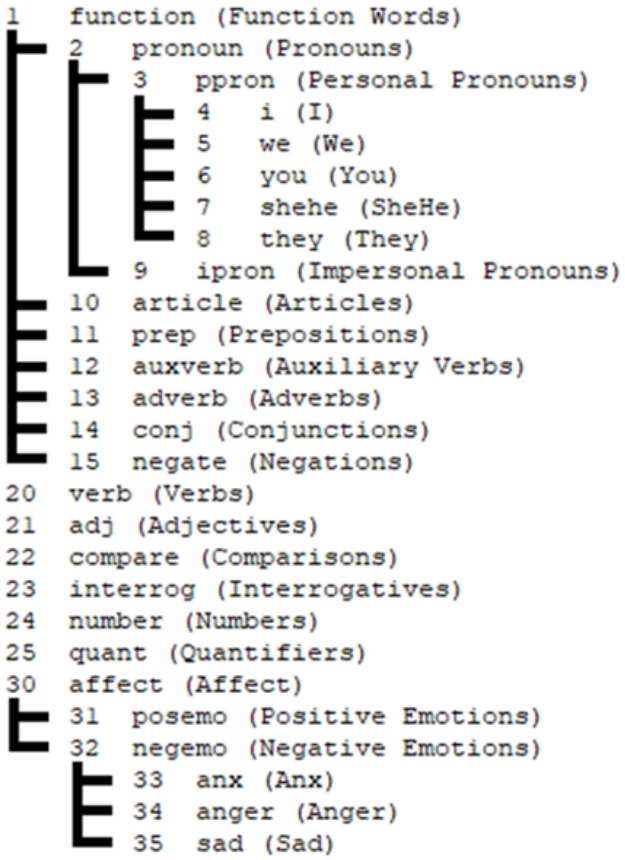

Figura 2 – Exemplo da divisão das categorias do dicionário LIWC_2015en em subcategorias.

A categoria que representa respostas emocionais ou afetivas é chamada ‘affect’ e tem como subcategorias emoções positivas (‘posemo’) e emoções negativas (‘negemo’), com esta última incluindo como subcategorias tristeza, ansiedade e raiva (respectivamente, ‘sad’, ‘anx’ e ‘anger’). Caso uma palavra esteja incluída em uma subcategoria, como a palavra ‘chorou’ que está associada à tristeza (‘sad’), ela também estará incluída nas categorias superiores. Além disso, a palavra pode estar incluída em outras categorias do dicionário, relacionadas à aspectos linguísticos ou sociais, conforme ilustrado na Figura 3.

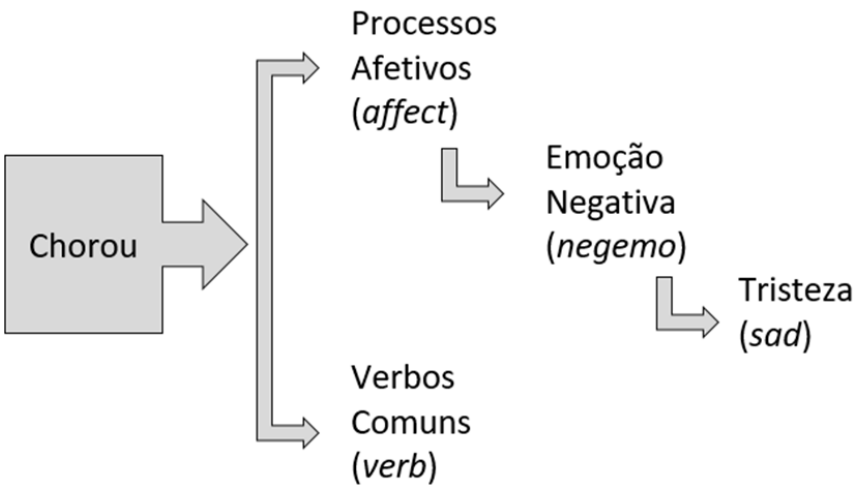

Figura 3 – Exemplo da associação de uma palavra (‘chorou’) a categorias (‘affect’, ‘verb’) e subcategorias (‘negemo’, ‘sad’) diversas.

A escolha das palavras para cada categoria do LIWC_2015en foi o resultado do desenvolvimento de muitas etapas ao longo de vários anos. Originalmente, a ideia era identificar um grupo de palavras mais estudadas em psicologia social, de saúde e de personalidade. Com o tempo, o domínio das categorias de palavras foi ampliado consideravelmente para além das dimensões emocionais e cognitivas básicas (PENNEBAKER et al., 2015b).

2.2- LIWC_2007pt / LIWC em português

O LIWC_2007pt é um dos sete dicionários afetivos disponível em português brasileiro, elaborados entre 2007 e 2016 (CRUZ et al., 2017). Se trata do resultado

do esforço colaborativo de três equipes: NILC[2], agência Checon Pesquisa e Unisinos[3]. Como representantes destas equipes, estão, respectivamente a tradutora Mônica Martins e pesquisadoras Rosangela Checon e Rove Chishman.

A tradução do dicionário de 2007 do inglês para o português foi feita usando vários Dicionários Bilíngues Português-Inglês. No processo de inclusão de 127.149 palavras em 64 categorias, levantadas automaticamente, as conjugações foram automaticamente incluídas usando o dicionário NILC Unitex-PB. Afirma-se que nenhuma revisão do trabalho manual da tradução foi feita, e também que pode ser melhorada[4].

É importante destacar que um grande número de palavras no LIWC_2007pt pode não representar um impacto positivo no número de palavras a serem contadas nos textos. No LIWC_2007pt, há palavras com problemas de ortografia, categorização, uso do caractere curinga '*' ou todos esses associados (CARVALHO; SANTOS; GUEDES, 2018). A seguir, serão detalhados esses problemas e trazidos alguns exemplos que podem ser encontrados.

Na análise preliminar da categoria de pronomes (*pronoun*) do LIWC_2007pt, verificamos que dos 128 itens que estão classificados como pronomes, mais de 40 parecem estar indevidamente incluídos na categoria (CARVALHO et al., 2018). Além disso, 8 palavras estão classificadas indevidamente na categoria de pronomes pessoais (*ppron*), do total de 54 palavras incluídas. Na categoria de pronomes impessoais, 49 das 88 palavras não deveriam estar associados a esta categoria, como as palavras 'ele' e 'ela', que são pronomes pessoais.

Atentar para essa grande diferença encontrada nestas categorias de palavras é relevante pois pronomes são importantes para diferenciar a maneira como as pessoas expressam o que sentem e como elas se conectam com outras pessoas (PENNEBAKER, 2011a). Isso significa que é importante que um dicionário que será usada em tarefas de classificação para análise de sentimentos também esteja devidamente ajustado nos aspectos linguísticos. Porém, as discrepâncias de categorização no LIWC_2007pt não estão restritas a essas categorias destacadas, e tampouco se apresentam como um único tipo de problema que encontramos.

Erros de ortografia também podem ser encontrados, inclusive juntos com inconsistências na categorização. No caso, encontramos cadastradas as palavras 'ninguén',

[2]Centro Interinstitucional de Linguística Computacional
[3]Universidade do Vale do Rio dos Sinos
[4]http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc, conforme acessado em 3 de maio de 2019.

categorizada apenas como um dos pronomes impessoais (*ipron*), e ‘ninguém’. A palavra sem erro ortográfico ‘ninguém’, por sua vez, não tinha a associação à categoria *ipron*. Removemos a palavra com erro de ortografia e fizemos a associação, na palavra com a grafia correta ‘ninguém’, à categoria *ipron*.

Outra questão relevante está relacionada às inconsistências com palavras usando o ‘*’. A categoria ‘affect’ do LIWC_2007pt tem 28.475 palavras, e 2.892 palavras atribuídas a essa categoria usam o ‘*’. No entanto, é possível encontrar a inclusão desnecessária de palavras com a mesma sequência de caracteres adjacentes antes do ‘*’, por exemplo, é possível encontrar ’falso’ e ’falsos’, que serão ignorados, pois o programa executa a contabilização de acordo com as categorias associadas à palavra ‘falso*’. É até mesmo possível encontrar várias palavras terminadas em ‘*’ que compartilham a mesma sequência de caracteres adjacentes no começo, como ‘abal*’, ‘abalad*’ ou ‘aborre*’, ‘aborrec*’, ‘aborrec*’, ‘aborrecid*’, ‘aborrecido*’ e muitos outros.

Há também o caso em que os problemas anteriormente indicados aparecem associados, i.e são encontradas palavras com inconsistências causadas pela múltipla inclusão de palavras com a mesma sequência de caracteres adjacentes antes de um ‘*’, erros de ortografia e categorização. No LIWC_2007pt, encontramos o caso da palavra ‘aborecidíssim*’, por exemplo, e se nota que além da falta de um ‘r’, esta palavra é incluída apenas na categoria ‘affect’, mas não nas subcategorias ‘negemo’ ou ‘anger’, tal como ocorre com ‘aborrec*’ e ‘aborrecid*’. Essas questões resultam em uma contagem menos precisa de palavras nas categorias a que elas deveriam estar associadas.

2.3- Versões do LIWC em outras línguas

Existem versões em outras línguas, além do original em inglês, disponíveis para o dicionário LIWC. Foram encontradas versões em línguas da Europa, como alemão (WOLF et al., 2008), catalão (MASSÓ et al., 2013), espanhol (RAMÍREZ-ESPARZA et al., 2007), francês (PIOLAT et al., 2011), holandês (ZIJLSTRA et al., 2004; BOOT; ZIJLSTRA; GEENEN, 2017; WISSEN; BOOT, 2017), entre outras. Além destas, que utilizam alfabeto latino, também conhecido como alfabeto romano, se encontram versões para línguas que usam outros sistemas de escrita como sérvio (BJEKIĆ et al., 2014) e russo (KAILER;

CHUNG, 2011), que usam o alfabeto cirílico, línguas como coreano (C. H. LEE; SHIM; YOON, 2005), além daquelas que utilizam um sistema de escrita logográfica, como Japonês (SHIBATA et al., 2016), Chinês Tradicional (HUANG et al., 2012) e Simplificado (GAO et al., 2013).

Nesses estudos se observa, conforme apontam (WISSEN; BOOT, 2017) e (BJEKIĆ et al., 2014), que a tarefa de desenvolver um dicionário do LIWC, a partir do dicionário em inglês, não é direta como a tradução de uma lista de palavras, indo além de simplesmente se associar palavras equivalentes da língua inglesa. Primeiramente, existe a complicação de que quem está traduzindo precisa verificar quais equivalentes se enquadram em quais categorias, observando que cada palavra pode ser inserida em várias delas (PENNEBAKER et al., 2015b). Isso pode levar, por exemplo, à tradução de uma palavra várias vezes, uma para cada categoria em que apareceu (WISSEN; BOOT, 2017).

Para o francês, o uso de ligações entre termos gerou uma dificuldade ante ao funcionamento do LIWC, que analisa uma sequência de caracteres, entre duas ausências de caracteres (os 'espaços'), como uma palavra (PIOLAT et al., 2011). Esta questão, assim como a expansão para as diferentes flexões de gênero, grau, conjugação, entre outras, resultou em um grande aumento do número de registros no arquivo de dicionário (PIOLAT et al., 2011; RAMÍREZ-ESPARZA et al., 2007). Ainda se observa a dificuldade de que, em alguns casos, é necessário procurar a tradução de palavras que correspondam a diferentes aspectos culturais. No caso da tradução holandesa de 2007, isso foi feito incluindo nomes de sindicatos holandeses na categoria relacionada a trabalho ('work') e bebidas holandesas na categoria relacionadas a lazer ('leisure') (WISSEN; BOOT, 2017).

A maioria das traduções encontradas foi realizada manualmente. Duas notáveis exceções foram as traduções para o catalão (MASSÓ et al., 2013) e para o holandês (WISSEN; BOOT, 2017), das versões de 2007 e 2015 do LIWC, respectivamente. Entretanto, cabe ressaltar que (WISSEN; BOOT, 2017) apontam que os autores do dicionário em catalão não relataram qualquer avaliação da tradução usando *corpus* paralelo.

Para o desenvolvimento do LIWC_2007pt, melhor detalhado na sub-seção 2.2, encontra-se a informação de que as conjugações foram inseridas automaticamente e as categorias de dicionário também foram levantadas automaticamente[5]. Conforme apontam (WISSEN; BOOT, 2017) em avaliações que realizaram, o processo que envolve tradução

[5]http://www.nilc.icmc.usp.br/portlex/index.php/pt/projetos/liwc, conforme acessado em 3 de maio de 2019.

automática ‘quase’ consegue um resultado tão bom quanto o processo manual, quando se olha para os coeficientes de correlação com o dicionário de referência (LIWC_2015en). Mesmo assim, reconhecem que uma correção manual posterior do dicionário traduzido automaticamente, apesar de melhorar os valores encontrados, ainda não os equipara aos valores que se obtêm por meio de uma tradução manual. Dessa forma, destacamos que o processo de tradução manual possui vantagens, o que reforça a escolha do processo que envolve a tradução manual para desenvolvimento tanto de inúmeras versões em outras línguas, quanto do LIWC_2015pt.

3- Metodologia

Neste capítulo, apresentamos como desenvolvemos o LIWC_2015pt em etapas, detalhando a criação de dicionários com categorias selecionadas do LIWC_2015en e, em seguida, a integração desses dicionários em um único arquivo. Primeiro, detalhamos na Seção 3.1 a criação de um dicionário de pronomes denominado PronounBP (CARVALHO et al., 2018). Em seguida, abordamos na Seção 3.2 a criação de um dicionário afetivo em português do Brasil chamado AffectPT-br (CARVALHO; SANTOS; GUEDES, 2018), baseado na tradução das palavras em categorias afetivas do LIWC_2015en. Na Seção 3.3 concluímos o LIWC_2015pt listando a criação e integração dos arquivos de dicionário contendo as demais categorias e subcategorias do LIWC_2015en.

3.1- PronounBP

Pronomes são importantes em tarefas de CA e AS (PENNEBAKER, 2011a; OFEK et al., 2015). Palavras dessa categoria servem para diferenciar a maneira como as pessoas expressam seus sentimentos e como elas se conectam com outras pessoas (PENNEBAKER, 2011a). Por isso, iniciamos o desenvolvimento do LIWC_2015pt com a criação de um novo dicionário de pronomes, denominado PronounBP (CARVALHO et al., 2018).

Criamos o PronounBP usando como base as palavras existentes nas categorias de pronomes do dicionário LIWC_2007pt e uma lista de pronomes que nós criamos e chamamos de `LDP1` (CARVALHO et al., 2018), contendo pronomes organizados conforme as subcategorias de 'pronoun' do LIWC_2015en. Em seguida, efetuamos três procedimentos: (i) remoção de alguns pronomes de categorias inadequadas; (ii) redesignação de algumas palavras do LIWC_2007pt nas categorias de pronomes; (iii) inclusão de pronomes não encontrados no LIWC_2007pt.

Para elaborar o PronounBP, utilizamos o LIWC_2007pt e os dicionários padrão do LIWC, tanto o LIWC_2015en (PENNEBAKER et al., 2015a) quanto a versão anterior

(PENNEBAKER; BOOTH; FRANCIS, 2007). Para criação da LDP1, organizamos um conjunto de recursos de referência da língua portuguesa, que chamamos de RLP (para referências no texto), contendo livros de referência gramatical (LUFT et al., 1997; DE ALMEIDA, 1973; TERRA; NICOLA, 2004), bem como o Dicionário Online Caldas Aulete[1] e Michaelis[2].

Também utilizamos o Vocabulário Ortográfico do Português (VOP), que funciona pela consulta por meio da Internet[3]. Se trata de uma ampla lista de palavras com designação das categorias morfossintáticas e das suas especificidades de flexão, e inclui as respectivas classificações gramaticais. O sistema de busca do VOP, na quinta edição (2009), permite a consulta a 381.000 palavras.

Para facilitar o processo de seleção de palavras e suas categorias, carregamos os dicionários do LIWC usando uma aplicação *web* desenvolvida em Java. Também implementamos códigos em R para o processo de comparação de listas de palavras propostas para inclusão no arquivo de dicionário.

3.1.1 Desenvolvimento do PronounBP

A metodologia que usamos para elaborar o PronounBP foi dividida em três etapas. Na primeira etapa, ilustrada na Figura 4, usamos o RLP (indicado na Seção 3.1) para criar a LDP1, com objetivo de incluir a maior quantidade possível de pronomes da língua portuguesa no novo dicionário. Após criada, verificamos a LDP1 com 3 pesquisadores: um psicólogo e dois linguistas.

Localizamos as palavras do LIWC_2007pt associadas às categorias de pronomes, conforme indicado no passo (1) da Figura 4. A LDP1 serviu para conferirmos se cada palavra estava corretamente associada a cada uma das categorias de pronomes, no passo (2) da Figura 4, para então incluirmos em uma versão inicial denominada PronounBP0. No caso de encontrarmos qualquer divergência em relação a LDP1, a associação da palavra à categoria de pronomes, assim como de suas subcategorias (se for o caso), seria removida, conforme indica o passo (3) da Figura 4.

[1] http://www.aulete.com.br/
[2] http://michaelis.uol.com.br/moderno-portugues/
[3] http://www.academia.org.br/nossa-lingua/busca-no-vocabulario

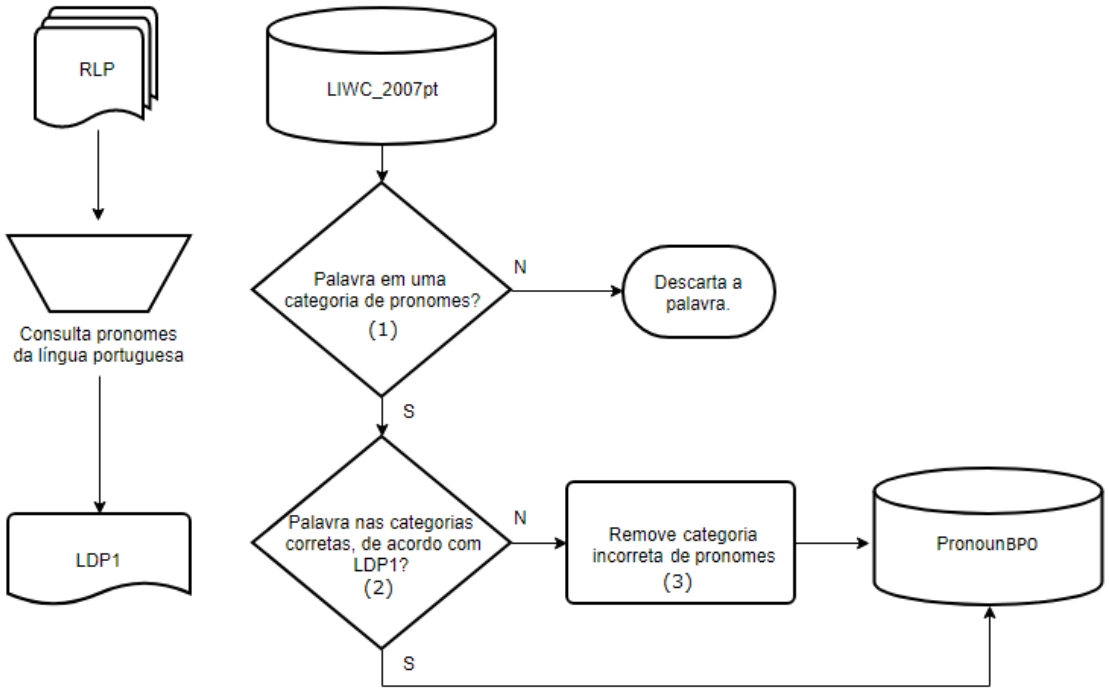

Figura 4 – Criação da LDP1 com uso do RLP e de uma versão inicial do dicionário de pronomes PronounBP0, a partir do LIWC_2007pt

Com esse processo anteriormente descrito, removemos do LIWC_2007pt todas as palavras da Tabela 1, das categorias nas quais elas foram incluídas indevidamente. Após essa etapa, passamos a contar no PronounBP0 com 84 palavras na categoria *pronoun*, já que 44 palavras foram removidas. Além disso, PronounBP0 conta com o seguinte número de palavras para as categorias modificadas: *ppron* (46), *we* (6), *you* (21), *shehe* (10) e *ipron* (39). É importante mencionar que não modificamos as categorias *i* e *they* nesta etapa.

Após a remoção das palavras indevidamente incluídas nas categorias de pronomes, realizamos então a redesignação de categorias em palavras do PronounBP0. Para cada palavra de LDP1, verificamos se ela poderia ser encontrada em PronounBP0. Caso positivo, procedemos com a redesignação para cada categoria de pronome correta. Ilustramos esse processo de criação do PronounBP na Figura 5. Nesta etapa ainda não incluímos palavras que não são encontradas no LIWC_2007pt. A Tabela 2 mostra uma lista de palavras do LIWC_2007pt redesignadas no dicionário PronounBP para as categorias de pronome conforme LDP1.

Em uma etapa final, adicionamos o restante das palavras da LDP1 na categoria de pronomes e suas subcategorias. Como resultado, o PronounBP apresenta mais palavras

Tabela 1 – Relação das palavras do LIWC_2007pt nas categorias relacionadas aos pronomes, com divergências em relação à adequação conforme LDP1

Nome da Categoria (*abbrev*): palavras	**Total**
Pronomes (*pronoun*): acontecimento*, algos, algures, aproximadamente, assim, assunto, assuntos, bens, bobagem, bobagens, bugiganga*, coisa, coisas, diferente, diferentes, fato, fatos, fêmea, idéia, idéias, issos, materiais, material, matéria, matérias, menina, moça, mulher, negócio, negócios, noção, noções, objeto, objetos, pertences, sozinha, sozinho, substância, substâncias, tolice*, traste*, troço, troços, tão;	44
Pronomes pessoais (*ppron*): bora, fêmea, menina, moça, mulher, sozinha, sozinho, vamos;	8
1ª pessoa do plural (*we*): bora, vamos;	2
2ª pessoa (*you*): a, as, o, as;	4
3ª pessoa singular (*shehe*): fêmea, homem, macho, menina, moça, mulher;	6
Pronomes impessoais (*ipron*): acontecimento, acontecimentos, algos, algures, alternada, alternadas, alternado, alternados, aproximadamente, assim, assunto, assuntos, bens, bobagem, bobagens, bugiganga*, coisa, coisas, diferente, diferentes, ela, ele, fato, fatos, idéia, idéias, lhe, materiais, material, matéria, matérias, negócio, negócios, ninguén, noção, noções, objeto, objetos, pertences, se, substância, substâncias, tolice, tolices, traste, trastes, troço, troços, tão;	49

na categoria *pronoun* do que LIWC_2007pt, conforme detalhado na Tabela 3. Enquanto a categoria de pronomes do LIWC_2007pt tem 128 palavras, essa categoria no PronounBP passou a ter 234.

Após a conclusão de elaboração do PronounBP, realizamos avaliações experimentais com o PronounBP e as categorias de pronomes encontradas no LIWC_2007pt. Nos testes, comparamos os resultados de classificação de autores a partir de textos, considerando a faixa etária. Os detalhes estão indicados nas seções do capítulo 4.

3.2- AffectPT-br

Criamos o AffectPT-br usando o LIWC_2015en como principal referência para construir a estrutura hierárquica de categorias e palavras que representam respostas emocionais ou afetivas a serem incluídas (CARVALHO; SANTOS; GUEDES, 2018). A principal categoria se chama ‘affect’, como no LIWC_2015en, e esta contém as subcategorias de emoções positivas (‘posemo’) e emoções negativas (‘negemo’). Também conta

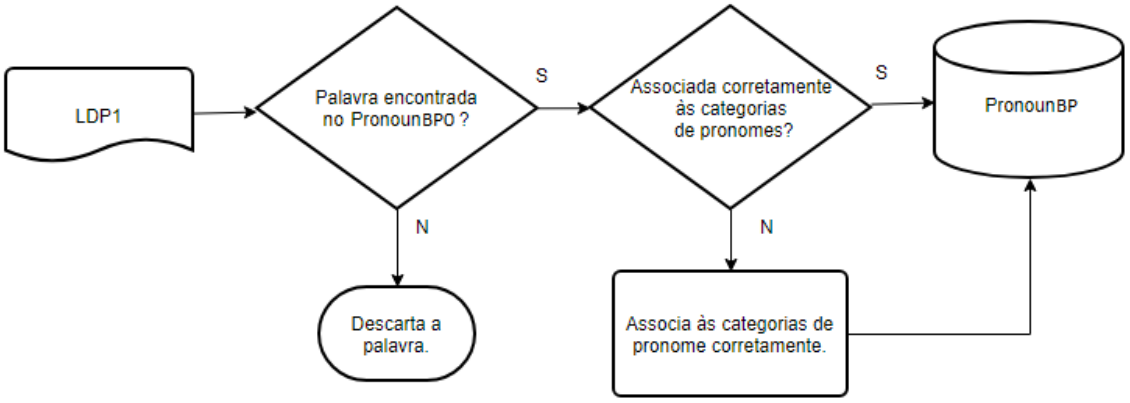

Figura 5 – Segunda etapa da criação do PronounBP, usando `LDP1`

Tabela 2 – Palavras no LIWC_2007pt redesignadas em categorias relacionadas a pronomes no PronounBP

Nome da categoria (abrev): palavras redesignadas	**Total**
Pronomes (*pronoun*): algum, alguma, algumas, alguns, bem, cada, certa, certas, certo, certos, consigo, dada, dado, dele, desse, desses, deste, destes, determinada, determinado, disso, mais, mesmas, mesmos, muita, muitas, muito, muitos, nada, nenhum, nenhuma, nenhumas, nenhuns, onde, picas, pouca*, poucas, pouco, poucos, quanto, semelhante, si, tanto, toda, todo, um, uma, umas, uns, várias, vários, vc, vcs;	53
Pronomes pessoais (*ppron*): consigo, ele, si, vc, vcs;	5
2ª pessoa (*you*): vc, vcs;	2
3ª pessoa singular (*shehe*): consigo, si;	2
Pronomes impessoais (*ipron*): algum, alguma, algumas, alguns, bem, cada, certa, certas, certo, certos, dada, dado, desse, desses, deste, destes, determinada, determinado, disso, mais, mesmas, mesmos, muitas, muito, muitos, nada, nenhum, nenhuma, nenhumas, nenhuns, ninguém, o, onde, picas, pouca*, poucas, pouco, poucos, quanto, semelhante, tanto, toda, todo, um, uma, umas, uns, várias, vários;	49

com as subcategorias de ‘negemo’ que estão associadas a tristeza, raiva e ansiedade: ‘sad’, ‘anger’ e ‘anx’, respectivamente.

3.2.1 Desenvolvimento do AffectPT-br

O desenvolvimento do AffectPT-br iniciou-se buscando opções de tradução das palavras que representam respostas emocionais ou afetivas encontradas no LIWC_2015en. Depois, os possíveis significados eram conferidos com auxílio de dicionários, sendo

Tabela 3 – Comparação da quantidade de palavras do LIWC_2007pt e PronounBP

Categoria	Abreviação	LIWC_2007pt	Removidas do LIWC_2007pt	PronounBP
Pronomes	pronoun	128	44	234
Pron. Pessoais	ppron	54	8	79
1ª pessoa do singular	i	7	-	9
1ª pessoa do plural	we	8	2	11
2ª pessoa	you	25	4	36
3ª pessoa do singular	shehe	16	6	17
3ª pessoa do plural	they	11	-	15
Pron. Impessoais	ipron	88	49	157

também verificado se a palavra equivalente em Português estava adequada, considerando a categoria sendo trabalhada. Para evitar duplicidade, também era verificado se a palavra equivalente em Português não estava incluída no AffectPT-br, sendo consultado um dicionário bilíngue para buscarmos sinônimos, conforme ilustramos na Figura 6.

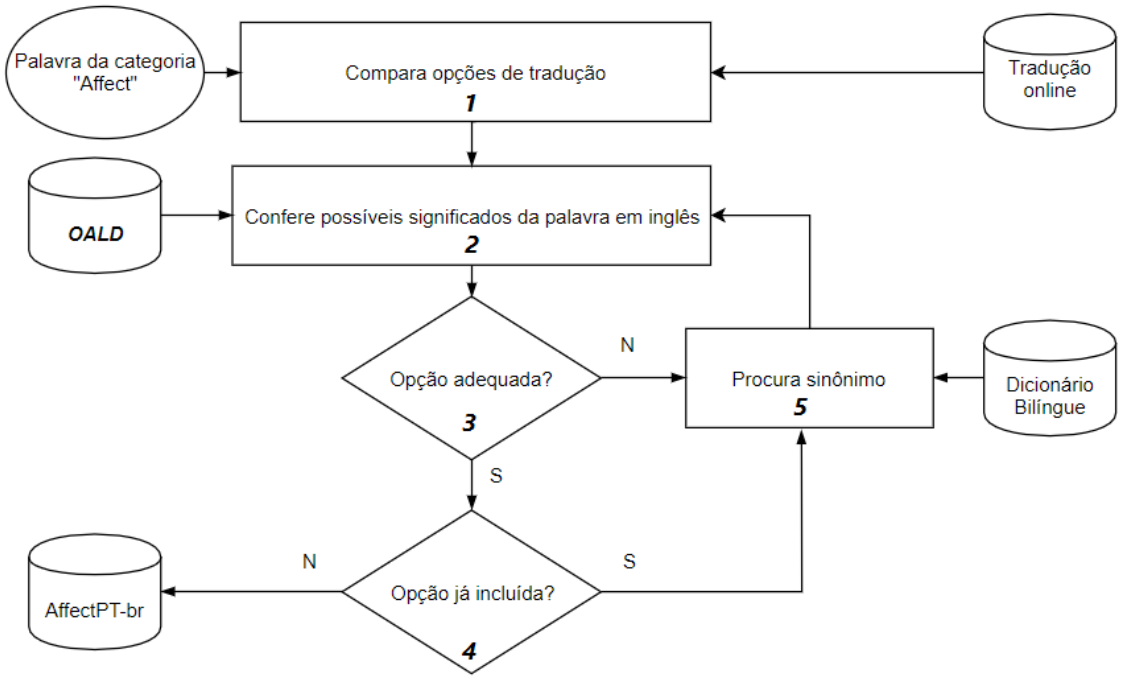

Figura 6 – Etapas do processo de desenvolvimento do AffectPT-br

Na primeira etapa do processo de desenvolvimento do AffectPT-br, extraímos do LIWC_2015en as subcategorias 'posemo' e 'negemo' e utilizamos serviços de Tradução *online*, enviando as palavras para o Google Tradutor (GT)[4] e o Bing Tradutor (BT)[5]. Tanto o

[4] https://translate.google.com/
[5] https://www.bing.com/translator

GT como o BT oferecem serviços que apresentam pequenas diferenças psicolinguísticas quando comparadas à tradução humana (RODRIGUES et al., 2017b). Entendemos este procedimento como sendo satisfatório nesta etapa, como indicado por (REIS et al., 2015).

Consideramos as possíveis traduções e escolhemos as palavras para inclusão nas subcategorias em português (1). Usamos o Dicionário *Oxford Advanced Learner's* (OALD) (HORNBY; WEHMEIER; ASHBY, 2000) tanto para uma melhor compreensão dos possíveis significados de cada palavra em inglês extraída do LIWC_2015en, quanto para decidir entre as opções do GT ou do BT (2). Caso a opção não estivesse adequada (3) e para evitar repetições (4), consultamos as palavras em inglês no Dicionário Bilíngue Inglês-Brasileiro Oxford (TEMPLE, 2001) para visualizar facilmente uma lista de sinônimos em português (5).

Como resultado, o AffectPT-br tem um total de 1.139 palavras atribuídas na categoria 'affect', 479 na subcategoria 'posemo' e 661 na subcategoria 'negemo'. Na Tabela 4, comparamos o número de palavras atribuídas a categorias e subcategorias afetivas do LIWC_2015en e do LIWC_2007pt com o AffectPT-br. O menor número de palavras de AffectPT-br se deve ao esforço de otimizar o uso do caractere curinga '*', evitando incluir todas as palavras com formas flexionadas para diferentes conjugações, varições de gênero, grau, etc.

Tabela 4 – Número de palavras em cada categoria afetiva do LIWC_2015en, LIWC_2007pt e AffectPT-br.

Category	LIWC_2015en	LIWC_2007pt	AffectPT-br
affect	1.393	28.475	1.139
posemo	620	12.878	479
negemo	744	15.115	661
anx	116	3012	107
anger	230	6867	188
sad	136	3864	135

3.3- LIWC_2015pt

Desta forma, contemplamos as duas abordagens distintas que escolhemos para criação do LIWC_2015pt e dos arquivos de dicionário. Repetimos o procedimento de criação do AffectPT-br e também fizemos uso de listas de palavras do domínio da ca-

tegoria trabalhada (como no caso da criação do PronounBP), e criamos dicionários contendo as categorias e subcategorias de Aspectos Sociais (‘social’), Processos Cognitivos (‘cogproc’), Processos Perceptivos (‘percept’), Processos Biológicos (‘bio’), Impulsos (‘drives’), Relatividade (‘relativ’) e Linguagem Informal (‘informal’). De forma semelhante, criamos dicionários contendo grupos de categorias relacionadas a Orientação Temporal (‘timeorient’) e Preocupações Pessoais (‘persconc’).

Criamos também dicionários contendo subcategorias de ‘Dimensões linguísticas’ segundo (PENNEBAKER et al., 2015a), que são relacionadas a Aspectos Gramaticais ou Semânticos (AGS). Estão nesse grupo as subcategorias da categoria Palavras de Função (‘funct’): Pronomes (‘pronoun’), Artigos (‘article’), Preposições (‘preps’), Verbos Auxiliares (‘auxverb’), Advérbios Comuns (‘adverb’) e Conjunções (‘conj’). Outros dicionários com categorias isoladas tais como Verbos (‘verb’), Quantificadores (‘quant’), entre outras, também estão relacionados a AGS e por isso pertencem a este agrupamento de ‘Dimensões linguísticas’.

PronounBP contém a totalidade de Pronomes (‘pronoun’), que é uma das subcategorias da categoria Palavras de Função (‘funct’). Por isso, decidimos unir este dicionário com os demais arquivos contendo subcategorias de Palavras de Função (‘funct’), começando pela subcategoria Artigos (‘article’). Com o arquivo que produzimos nessa união, repetimos o processo com o dicionário de subcategorias de Preposições, Verbos auxiliares, Advérbios Comuns, Conjunções e Negações, obtendo assim um dicionário com Palavras de Função (‘funct’) e todas as suas subcategorias.

Utilizando processo semelhantemente de união de dicionários, produzimos um dicionário contendo tanto as Palavras de Função (‘funct’) como as demais categorias de AGS, i.e. Verbos (‘verb’), Adjetivos (‘adj’), Comparações (‘compare’), Interrogativas (‘interrog’), Números (‘number’) e Quantificadores ‘quant’. É importante observar que no dicionário com estas categorias, as variações referentes às formas flexionadas de palavras da categoria ‘adj’ (e.g. ‘calma’, ‘calmas’, ‘calmo’, ‘calmos’) estão incluídas, assim como a forma na categoria ‘adverb’ (e.g. ‘calmamente’).

Isso é relevante, pois na integração do dicionário com palavras AffectPT-br precisamos diminuir o uso das palavras terminadas com ‘*’ (asterisco). Isso é necessário para não haver inconsistências com as variações de gênero, número e grau de diferentes palavras nas categorias de AGS, como ‘adj’ e ‘adverb’. Também incluímos as formas conjugadas dos verbos e atribuímos a categorização de acordo com o foco no presente

('focuspresent'), passado ('focuspast') e no futuro ('focusfuture').

Finalizamos o LIWC_2015pt repetindo, por etapas, a união de cada dicionário produzido por esse processo com os dicionários contendo as categorias e subcategorias de 'social', 'cogproc', 'percept', 'relativ' e 'informal', e também com o dicionário que contém as categorias agrupadas em 'persconc'. A quantidade de palavras nas principais categorias do LIWC_2015pt, assim como as quantidades no LIWC_2015en e no LIWC_2007pt podem ser observadas e comparadas na Tabela 5. No Apêndice A, incluímos todas as categorias e subcategorias dos dicionários para comparação. No total, o LIWC_2015pt que apresentamos nesse trabalho possui um total de 14.459 palavras em 73 categorias.

Tabela 5 – Comparação da quantidade de palavras das principais categorias do LIWC_2015en (2015), LIWC_2007pt (2007_pt) e LIWC_2015pt (2015_pt)

Categoria	Abreviação	Exemplos	2015	2007_pt	2015_pt
Palavras de função	funct	Para, não, muito	491	5.512	1.426
Verbos	verb	Conquistar, superar	1.000	23.873	6.162
Adjetivos	adj	Incrível	764	NA	1.241
Comparações	compare	Antes, maior	317	NA	300
Interrogativas	interrog	Aonde, como	48	NA	23
Números	number	Mil, três	36	83	106
Quantificadores	quant	Alguns, pouco	77	622	160
Proc. afetivos	affect	Admirável, agonia	1.393	28.475	2.105
Social	social	Amiga, eles	756	13.634	1.445
Proc. cognitivos	cogproc	Devia, porque, sabe	797	46.308	2.691
Proc. perceptivos	percept	Ouvir, ver, sentir	436	17.607	1.001
Proc. Biológicos	bio	Comeu, dor, sangue	748	17.861	1.924
Impulsos	drives	Ambiciosa, clã	1.103	NA	3.176
Foco passado	focuspast	Falou, passado	341	7.684	3.159
Foco presente	focuspresent	Agora, atual	424	4.715	1.416
Foco futuro	focusfuture	Amanhã, promessa	97	268	1.456
Relatividade	relativ	Continuar, indo	974	24.966	3.220
Trabalhos	work	Assessoria, baia	444	7.735	781
Lazer	leisure	Baile, pousada	296	6.331	556
Casa	home	Aluguel, quintal	100	2.019	145
Dinheiro	money	Ações, pensão	226	5.353	356
Religião	relig	Fé, santo, zen	174	2.066	343
Morte	death	Funeral, morreu	74	2.429	90
Linguagem informal	informal	Buguei, td, vcs	380	NA	419

4- Análise

O objetivo nesta seção é estabelecer comparações entre o LIWC_2007pt e o LIWC_2015pt por meio dos experimentos apresentados, além de apresentar comparações com foco em importantes categorias do dicionário desenvolvido: Pronomes (‘pronoun’) e Processos Afetivos (‘affect’). Abordaremos na Seção 4.1 o desempenho em tarefas de classificação, típica da área de AS, referente à aplicação de algoritmos de classificação na tarefa de inferência de faixa etária e classificação de polaridade de emoções. Realizamos na Seção 4.2 a comparação estatística dos resultados obtidos a partir de uma amostra de textos bilíngues. Ao comparar o coeficiente de correlação entre análises realizadas com o LIWC_2015en e o LIWC_2015pt, com o coeficiente de correlação de análises com o LIWC_2015en e o LIWC_2007pt, buscamos uma avaliação da curadoria mais detalhada, observando os valores para cada categoria dos dicionários.

Em cada experimento, configuramos o programa LIWC para utilizar o dicionário que queremos avaliar. Analisamos todas as postagens do conjunto de dados escolhido com LIWC, que processa o texto das postagens e gera um arquivo em que registra os valores percentuais de palavras de cada categoria encontradas no texto.

4.1- Comparação dos resultados em tarefas de classificação

4.1.1 Classificação para inferência da faixa etária

Conforme mencionado anteriormente, o principal objetivo desse trabalho consiste em melhorar a classificação de textos em português em tarefas de MT utilizando o LIWC, por meio do desenvolvimento de um dicionário em português para uso na versão 2015 do LIWC. Para realizar a comparação dos resultados da tarefa de classificação da inferência da faixa etária, coletamos dados de uma rede social brasileira chamada Meu Querido

Diário (MQD)[1]. Escolhemos o MQD porque os usuários escrevem predominantemente em português do Brasil. O conjunto de dados, chamado `MQD190k`, consiste em 190.000 entradas (CARVALHO et al., 2018). Está dividido em 3 classes, como em (SCHLER et al., 2006): 10s, 20s e 30s, contendo usuários de ambos os gêneros com idades entre 13 e 17, 23 a 27 e 33 a 42 anos, respectivamente. O intervalo de 6 anos entre o final de uma faixa e início de outra foi planejado para uma diferenciação mais clara.

Considerando a alteração percebida no uso de pronomes ao longo da vida (SCHLER et al., 2006), produzimos um experimento para classificar a faixa etária a partir de textos, usando como dicionário o PronounBP (CARVALHO et al., 2018), que corresponde às categorias de pronomes que incluímos no LIWC_2015pt. Após o processamento dos textos com o PronounBP, os arquivos gerados pelo LIWC são utilizados na tarefa de classificação, usando os algoritmos: NB e NBM, por serem métodos de base de referência para classificação de texto (WANG; MANNING, 2012), e DT e J48, por fornecerem bons resultados na classificação de textos (FERSINI; POZZI; MESSINA, 2015; GABRILOVICH; MARKOVITCH, 2004). Outro algoritmo escolhido foi o LMT, por ser um dos que melhor funciona para classificar textos utilizando recursos estilísticos do português (AIRES et al., 2004), apresentando inclusive bons resultados em instâncias com o LIWC como um dos recursos utilizados em algumas tarefas (e.g., detecção de sátira, detecção de sarcasmo) (K. RAVI; V. RAVI, 2017).

Apresentamos na Tabela 6 os resultados de F_1 usando os algoritmos NB, NBM, DT, LMT e J48 para classificação da inferência da faixa etária dos usuários, utilizando exclusivamente a categoria de pronomes e suas subcategorias. Pode-se notar que os algoritmos NB, NBM, DT, LMT e J48 apresentam um desempenho melhor com o uso de arquivos processados com o uso do dicionário PronounBP do que com o LIWC_2007pt. O valor de F_1 usando o algoritmo NBM foi o que apresentou melhor resultado, tendo o uso do PronounBP ocasionado um incremento de até $1,2\%$ em relação ao uso do LIWC_2007pt.

Tabela 6 – Valor de F_1 dos algoritmos de classificação da inferência da faixa etária dos usuários do MQD, usando exclusivamente a categoria de pronomes e suas subcategorias.

	NB	NBM	DT	LMT	J48
PronounBP	**0,508**	**0,512**	**0,498**	**0,503**	**0,507**
LIWC_2007pt	0,496	0,506	0,491	0,500	0,503

[1] http://www.meuqueridodiario.com.br

Conduzimos então um segundo experimento de classificação usando os valores analisados de todas as 73 categorias do LIWC_2015pt e 64 do LIWC_2007pt. Conforme pode ser observado na Tabela 7, usamos quatro dos cinco algoritmos de classificação no primeiro experimento, sendo que substituímos o algoritmo DT considerando que este apresenta o pior desempenho. No lugar, usamos o algoritmo RF, considerando que é um classificador que apresenta bons resultados em tarefas de classificação de textos em português (CARVALHO; SANTOS; GUEDES, 2018; NASCIMENTO; DUARTE; GUEDES, 2018). Observamos na Tabela 7 que o algoritmo RF apresenta resultados melhores que os demais algoritmos, tanto com a utilização de dados do processamento com o LIWC_2015pt quanto com o LIWC_2007pt.

Para verificar se as diferenças encontradas são significativas estatisticamente, realizamos testes t de (STUDENT, 1908) para dados pareados (valores de F_1 dos algoritmos para cada dicionário), considerando 4 graus de liberdade (DF=4)[2]. A hipótese nula é que os valores são iguais, i.e. não há diferença (significativa) entre o uso do LIWC_2007pt ou LIWC_2015pt. A hipótese alternativa, em que o valor da probabilidade p é menor que o valor especificado, assume a diferença entre uso do LIWC_2007pt e LIWC_2015pt como significativas estatisticamente.

A utilização do teste t pressupõe que os valores tem distribuição normal (STUDENT, 1908), por se tratar de um teste paramétrico (CROUX; DEHON, 2010). Por isso, testamos neste caso e nos demais apresentados adiante, os valores conferidos usando o teste de Shapiro-Wilk (S-W) (SHAPIRO; WILK, 1965). Verificamos que esse teste é indicado como uma das melhores escolhas para testar a normalidade dos dados (THODE, 2002).

Tabela 7 – Valor de F_1 dos algoritmos de classificação da inferência da faixa etária dos usuários do MQD, usando todas as 73 categoria do LIWC_2015pt e 64 LIWC_2007pt

	NB	NBM	RF	LMT	J48
LIWC_2015pt	**0,538**	**0,543**	**0,572**	**0,570**	**0,528**
LIWC_2007pt	0,536	0,527	0,566	0,563	0,521

Na comparação dos resultados de classificação para inferência da faixa etária, os testes t pareados mostraram que as melhorias são estatisticamente significativas no nível de confiança de 95% em ambos os experimentos. Na comparação com as categorias de pronomes, calculamos que o valor de t é $4,08$ e o valor de p é $0,015$, por isso rejeitamos a

[2]De forma a termos medidas mais robustas, seria necessário observar o número de instâncias que levaram aos valores de F_1 dos algoritmos de classificação.

hipótese nula considerando que encontramos $p < 0,05$. O tamanho do efeito padronizado observado é grande $(1,82)$. Na comparação com todas as categorias do LIWC_2015pt e LIWC_2007pt, observamos o tamanho do efeito padronizado igual a $1,48$, sendo que o valor encontrado de t é $3,31$ e o valor de p é $0,029$, i.e. resultado mostra que a diferença é significativa, pois $p < 0,05$.

4.1.2 Classificação de polaridade de emoções

Para a avaliação do dicionário LIWC_2015pt na classificação de polaridade de emoções, conduzimos duas sequências de experimentos. Em cada uma, carregamos no LIWC os dicionários LIWC_2015pt e LIWC_2007pt e analisamos dois conjuntos de dados de redes sociais, usando (i) todas as categorias e (ii) somente a categoria de afeto ('affect') e suas subcategorias. O objetivo principal é comparar o desempenho entre as categorias afetivas do LIWC_2015pt e do LIWC_2007pt.

Um dos conjuntos de dados utilizado é o `MQD60k`, que contém dados coletados do MQD (CARVALHO; SANTOS; GUEDES, 2018). Além da disponibilidade de obtenção destes dados[3], outro fator de escolha deste conjunto se deve à associação que é feita pelo usuário/autor de cada publicação a uma emoção. Cada emoção é uma das seis emoções propostas por Paul Ekman (i.e. raiva, felicidade, tristeza, surpresa, medo e nojo) (EKMAN, 1992). No `MQD60k`, há um total de 59.166 posts com emoções de felicidade e tristeza associadas, selecionados aleatoriamente de usuários na faixa entre 13 e 99, sem distinção de gênero. Assim, este conjunto de dados contém 32.244 posts associados à classe felicidade e 26.922 com classe tristeza.

O segundo conjunto de dados que usamos é o `TAS-PT`, também disponível publicamente[4], composto de dados coletados do Twitter[5] (CAVALCANTE; MALHEIROS, 2017). O conjunto de dados `TAS-PT` contém dois arquivos: (i) um arquivo com códigos de identificação de publicações rotuladas para 'sentimentos positivos'; (ii) um arquivo com códigos de identificação de publicações rotuladas para 'sentimentos negativos'.

[3] `MQD60k` pode ser obtido em `https://github.com/LaCAfe/MQD60k`

[4] `TAS-PT` pode ser obtido em `https://github.com/pauloemmilio/dataset`

[5] Em 2019, o Twitter é uma rede social que permite aos usuários troca de mensagens com outros usuários, limitada a textos de até 280 caracteres. As publicações de um usuário são exibidas em seu perfil e enviadas também a outros usuários que tenham optado em acompanhar suas publicações.

Como nenhum deles tinha o conteúdo textual das publicações, usamos a Interface de Programação de Aplicação (API) do Twitter para recuperar o conteúdo das mensagens por meio dos códigos de identificação nos arquivos. O processo resultou na criação do conjunto de dados `TAS-PT-60k` (CARVALHO; SANTOS; GUEDES, 2018), com 59.260 de conteúdo textual de publicações da `TAS-PT`, em que 28.853 são negativos e 30.407 são positivos.

A Tabela 8 mostra os resultados com os valores de F_1 usando somente com o uso das categorias de afeto do LIWC_2015pt e do LIWC_2007pt para análise do conjunto de dados `MQD60k`. Os melhores valores de F_1 são obtidos utilizando dados obtidos pelo processamento dos textos com as categorias de afeto do LIWC_2015pt, sendo possível notar que o algoritmo LMT produz o melhor valor de F_1 $(0,722)$ neste conjunto de dados. O algoritmo LMT também produz o melhor valor de F_1 $(0,687)$ com os dados obtidos pelo processamento dos textos com as categorias de afeto do LIWC_2007pt.

Tabela 8 – Valor de F_1 dos algoritmos usados na classificação de polaridade de emoções para o conjunto de dados `MQD60k`, usando somente as categorias de afeto.

	NB	NBM	RF	LMT	J48
LIWC_2015pt	**0,683**	**0,715**	**0,701**	**0,722**	**0,719**
LIWC_2007pt	0,678	0,675	0,668	0,687	0,683

Na Tabela 9, temos o resultado do experimento de classificação usando os valores analisados de todas as 73 categorias do LIWC_2015pt e 64 do LIWC_2007pt. Usando mais categorias, o algoritmo RF apresenta melhor resultado na utilização de dados obtidos pelo processamento dos textos com o LIWC_2015pt. Com o LIWC_2007pt os algoritmos LMT e RF apresentaram melhores resultados.

Tabela 9 – Valor de F_1 dos algoritmos usados na classificação de polaridade de emoções para o conjunto de dados `MQD60k`, usando todas as categorias dos dicionários.

	NB	NBM	RF	LMT	J48
LIWC_2015pt	**0,637**	**0,725**	**0,777**	**0,775**	**0,710**
LIWC_2007pt	0,631	0,711	0,752	0,753	0,682

Testes t pareados mostraram que as melhorias são estatisticamente significativas no nível de confiança de 95% e DF=4. Na comparação com as categorias de afetivas, o valor de t é $4,73$ e o valor de p é $0,009$, i.e. resultado é significativo para $p < 0,05$. Na comparação com todas as categorias do LIWC_2015pt e LIWC_2007pt, o valor de t é $4,75$ e o valor de p é $0,009$, i.e. resultado é significativo para $p < 0,05$. Em ambos os casos,

observamos que o tamanho de efeito padronizado é grande, sendo equivalente a $2,11$ e $2,12$, respectivamente.

Os resultados para o conjunto de dados `TAS-PT-60k` são descritos na Tabela 10. Os melhores resultados utilizando tanto o LIWC_2007pt quanto o LIWC_2015pt foram alcançados com o algoritmo LMT, enquanto que, usando o LIWC_2015pt, os classificadores RF e J48 também alcançaram bons resultados. Com isso, observamos que o LIWC_2015pt, usando somente as categorias afetivas, atinge no total uma melhoria de até $34,1$ % sobre o valor de $0,687$, o melhor resultado com o LIWC_2007pt e suas categorias afetivas, no mesmo conjunto de dados.

Tabela 10 – Valor de F_1 dos algoritmos usados na classificação de polaridade de emoções para o conjunto de dados `TAS-PT-60k`, usando somente as categorias de afeto.

	NB	NBM	RF	LMT	J48
LIWC_2015pt	**0,799**	**0,889**	**0,920**	**0,921**	**0,920**
LIWC_2007pt	0,671	0,675	0,668	0,687	0,683

Na Tabela 11, temos o resultado do experimento de classificação usando os valores analisados de todas as 73 categorias do LIWC_2015pt e 64 do LIWC_2007pt. Usando mais categorias, o algoritmo RF apresenta melhores resultados utilizando dados do LIWC_2015pt e do LIWC_2007pt. Também foram realizados testes paramétricos comparando os valores dos classificadores para verificar se existe diferença significativa pelo uso de cada um dos dicionários.

Tabela 11 – Valor de F_1 dos algoritmos usados na classificação de polaridade de emoções para o conjunto de dados `TAS-PT-60k`, usando todas as categorias dos dicionários LIWC_2007pt e LIWC_2015pt.

	NB	NBM	RF	LMT	J48
LIWC_2015pt	**0,743**	**0,875**	**0,955**	**0,965**	**0,949**
LIWC_2007pt	0,615	0,649	0,697	0,683	0,644

Testes t pareados mostraram que as melhorias são estatisticamente significativas no nível de confiança de 95% (DF=4). Na comparação com as categorias de afetivas, o valor de t é $9,64$ e o valor de p é $0,001$, i.e. resultado é significativo para $p < 0,05$. Na comparação com todas as categorias do LIWC_2015pt e LIWC_2007pt, o valor de t é $7,77$ e o valor de p é $0,001$. O tamanho de efeito padronizado em cada um destes casos é grande, sendo equivalente a $4,31$ e $3,47$, respectivamente.

4.2- Comparação estatística

4.2.1 Conjunto de textos para análise com *Corpus* paralelo

Conduzimos um procedimento de avaliação semelhante ao realizado por (WISSEN; BOOT, 2017), (BJEKIĆ et al., 2014) , (PIOLAT et al., 2011), (WOLF et al., 2008) e (RAMÍREZ-ESPARZA et al., 2007), como uma etapa adicional mais detalhada de avaliação da curadoria. De forma a obtermos medidas para cada categoria dos dicionários e medir a correlação entre as versões, buscamos textos em português para serem analisados no LIWC, utilizando o LIWC_2015pt e o LIWC_2007pt como dicionários, e seus equivalentes em inglês utilizando o LIWC_2015en. Usamos um *corpus* que é uma coleção bilíngüe Português-Inglês das edições on-line da revista científica REVISTA PESQUISA FAPESP[6] (AZIZ; SPECIA, 2011), e chamamos de `FAPESP-CORPUS` nesse trabalho.

4.2.2 Cálculos

Submetemos cada elemento do `FAPESP-CORPUS` à análise pelo LIWC. Usamos o LIWC_2015pt e o LIWC_2007pt para analisar os textos em português, e o LIWC_2015en para os textos em inglês. O resultado dessa análise são tabelas contendo valores das frequências relativas de palavras nas categorias dos dicionários nas colunas, para cada arquivo analisado.

Utilizamos o teste de d'Agostino-Pearson (A-P) (D'AGOSTINO; BELANGER; D'AGOSTINO JR, 1990), para testar a hipótese nula: "Os valores percentuais no texto, de palavras na categoria do LIWC, podem ser modelados de acordo com a distribuição normal?". Na Tabela 12 estão disponíveis os valores de p encontrados para as categorias de pronomes, sendo que não encontramos valor de p no teste A-P alto o suficiente ($>0,01$) para aceitar que o conjunto de dados para estas categorias segue a distribuição normal.

Repetimos este teste com as demais categorias, encontrando resultados seme-

[6] `http://revistapesquisa.fapesp.br/`

lhantes, i.e. nenhuma das categorias apresenta valor de p no teste A-P alto o suficiente ($>0,01$). Desta forma, rejeitamos a hipótese nula de acordo com os valores de p calculados no teste A-P e adotamos a hipótese alternativa em que “os dados não seguem uma distribuição normal”. Isso se mostra relevante na definição da comparação a ser realizada pois, por definição, variáveis que não tenham distribuição normal e homogeneidade de variâncias não devem utilizar testes paramétricos (CROUX; DEHON, 2010). Considerando isso, escolhemos o coeficiente de correlação τ b de Kendall (KENDALL, 1938), que avalia as associações estatísticas e não depende de suposições sobre as distribuições (NOETHER, 1981).

Tabela 12 – Resultados do teste A-P com os valores de p para saber se a distribuição dos dados obtidos para cada categoria do LIWC poderiam ser modelados de acordo com a distribuição normal

	pronoun	ppron	i	we	you	shehe	they	ipron
A-P	$<$0,001	$<$0,001	$<$0,001	$<$0,001	$<$0,001	$<$0,001	$<$0,001	$<$0,001

4.2.3 Correlação com o LIWC

A comparação dos coeficientes de correlação τ b de Kendall dos resultados das categorias, tanto do dicionário padrão do LIWC 2015 em inglês com LIWC_2015pt, quanto do LIWC 2015 em inglês com LIWC_2007pt, ajudou a avaliarmos o desenvolvimento deste trabalho. Os valores dos coeficientes de correlação τ b de Kendall para as principais categorias do LIWC_2015pt podem ser observados na Tabela 13, sendo que incluímos no Apêndice B os valores obtidos com todas as categorias e subcategorias dos dicionários para comparação.

Para uma comparação estatística, utilizamos o teste dos postos sinalizados de Wilcoxon (GEHAN, 1965), considerado como uma versão não paramétrica do teste t pareado (KERBY, 2014), e calculamos o valor W e o valor z. Neste teste, se considera que a distribuição da estatística de Wilcoxon W tende a formar uma distribuição normal quando há pelo menos 20 elementos, e desta forma o valor z pode ser usado para avaliar a hipótese. Entretanto, se o tamanho de N for baixo, principalmente se estiver abaixo de 10, devemos utilizar o valor W. Partimos da hipótese nula “as medianas das duas

amostras são idênticas" como forma de expressar que "analisando textos em português com o LIWC_2015pt, em comparação com o LIWC_2007pt, não se observa diferença significativa no teste de correlação com os valores de análise de textos equivalentes em inglês com o LIWC_2015en".

Como o número de valores de τ analisados é maior do que 20, usamos valor z para avaliar a hipótese nula. O valor de z encontrado é 6.48, e desta forma verificamos valor $p < 0,00001$. Com isso, rejeitamos a hipótese nula e consideramos a alternativa de que "se observa diferença significativa no teste de correlação com a versão em inglês do LIWC".

Tabela 13 – Resultados da análise do FAPESP-CORPUS para comparação entre os valores da Mediana, Mínimo (Mín) e Máximo (Máx) das porcentagens das principais categorias do LIWC_2015en (En) e do LIWC_2015pt (Pt), e comparações entre LIWC_2015en x LIWC_2015pt (τ_1) e LIWC_2015en x LIWC_2007pt (τ_2), observando os coeficientes de correlação τ b de Kendall nos valores das categorias

Categoria		En			Pt			
(Abreviação)	Mediana	Mín	Máx	Mediana	Mín	Máx	τ_1	τ_2
funct	46,90	9,52	57,75	46,81	10,74	58,25	**0,54**	0,46
verb	9,00	0,00	17,23	7,92	0,00	17,16	0,46	**0,48**
adj	3,95	0,00	9,77	3,91	1,20	8,89	**0,47**	NA
compare	2,14	0,00	6,64	2,53	0,00	6,02	**0,56**	NA
interrog	1,19	0,00	5,69	3,79	0,00	11,00	**0,43**	NA
number	2,58	0,00	58,88	4,55	0,00	56,67	0,68	**0,73**
quant	1,74	0,00	5,47	2,01	0,00	6,76	**0,55**	0,33
affect	2,54	0,00	9,27	2,60	0,00	9,79	**0,59**	0,46
social	4,26	0,00	15,34	3,93	0,00	13,06	**0,60**	0,37
cogproc	8,21	0,93	17,49	10,09	0,47	21,62	**0,60**	0,44
percept	1,24	0,00	11,02	1,83	0,00	8,85	0,40	**0,41**
bio	1,10	0,00	11,32	1,64	0,00	10,77	**0,64**	0,47
drives	5,66	0,00	14,84	7,09	0,00	15,76	**0,59**	NA
focuspast	2,72	0,00	9,80	1,38	0,00	8,21	**0,63**	0,56
focuspresent	5,55	0,00	12,78	5,22	0,00	10,47	0,52	**0,54**
focusfuture	0,54	0,00	4,39	0,32	0,00	3,96	**0,48**	0,18
relativ	12,88	0,79	24,85	12,45	3,29	25,35	**0,57**	0,45
work	5,36	0,47	20,42	4,23	0,00	16,19	**0,77**	0,73
leisure	0,38	0,00	8,24	0,41	0,00	8,70	**0,44**	0,33
home	0,08	0,00	4,97	0,08	0,00	4,58	**0,49**	0,35
money	0,50	0,00	9,12	0,71	0,00	10,76	**0,60**	0,50
relig	0,03	0,00	7,24	0,06	0,00	6,32	**0,38**	0,36
death	0,00	0,00	3,24	0,00	0,00	3,26	**0,72**	0,41
informal	0,20	0,00	2,79	0,25	0,00	1,97	**0,01**	NA

4.2.4 Tempo de processamento

Além das medidas apresentadas, anotamos também o tempo decorrido (*Elapsed time*) (KARP; FLATT, 1990) de cada conjunto de texto que utilizamos. A Figura 7 apresenta os valores em milissegundos (ms) de execução do LIWC com os dicionários LIWC_2007pt e LIWC_2015pt. Os valores foram obtidos de um ambiente com processador Intel Core i3-330M com 2 núcleos de 2,13 GHz, memória RAM de 4,00 GB DDR3, placa-mãe modelo Calpella CRB, disco rígido de 5400 RPM modelo WDC WD5000BEVT-00A0RT0 em barramento ATA e Microsoft Windows 10 Professional 64-bit (Build 17134). Nessa medida, é considerado tanto o tempo de CPU quanto o tempo de espera do sistema (CROWL, 1994).

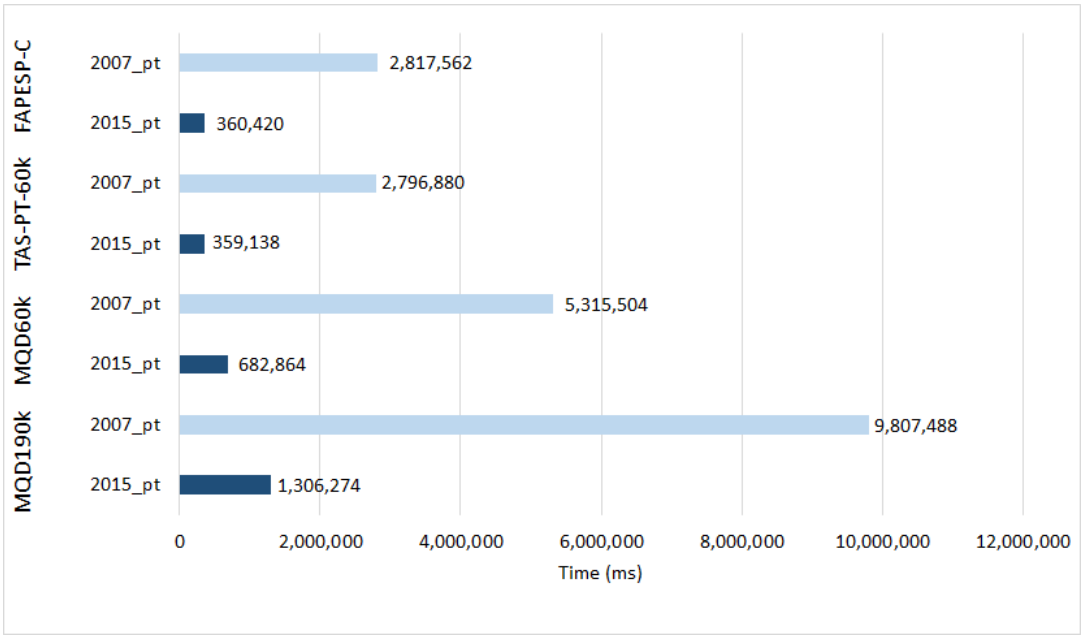

Figura 7 – Comparação do tempo (ms) de processamento dos diferentes conjuntos de textos, usando LIWC_2007pt e LIWC_2015pt.

Observamos que os valores utilizando o LIWC_2015pt reduziram em até 87% o tempo necessário para processamento utilizando o LIWC_2007pt. Usando o LIWC_2015pt, os tempos de processamento de `MQD190k`, `MQD60k` , `TAS-PT-60k` e `FAPESP-CORPUS` corresponderam respectivamente a $13,32\%$, $12,85\%$, $12,84\%$ e 12.79%.

Observamos que o tempo necessário para processar textos usando o LIWC_2015pt é menor que o tempo necessário para processar textos usando o LIWC_2007pt. Isto foi

possível considerando o menor número de palavras no dicionário, resultado do refinamento na escolha de palavras adequadas a cada categoria e da otimização do uso do caractere curinga '*'. Usando o LIWC_2015pt, cada tempo de processamento foi reduzido para 87% do tempo necessário para processamento usando LIWC_2007pt.

Considerações finais

Contribuições

O presente estudo teve como foco a melhoraria nos resultados em tarefas de classificação usando arquivos de textos processados pelo LIWC. Aproveitamos, na construção do LIWC_2015pt, anos de estudo exploratório de componentes emocionais, cognitivos e estruturais ao incorporar palavras traduzidas do LIWC_2015en. Desta maneira, buscamos a melhora de resultados em tarefas usadas na área de MT em Português do Brasil. Desenvolvemos este recurso para ser usado em estudos acadêmicos que extraem conhecimento a partir de fontes não-estruturadas, envolvendo conhecimentos da área de Computação Afetiva (CA) e Análise de Sentimentos (AS).

Dentre os desafios no desenvolvimento deste novo dicionário, destacamos a quantidade limitada de documentação sobre o dicionário em português na versão 2007 para uso nesse programa. Contribuímos no presente trabalho concentrando as informações encontradas sobre essa versão. Descrevemos alguns dos problemas existentes e comparamos resultados obtidos com esse dicionário.

Como consequência dos estudos realizados neste trabalho, ocorreu a divulgação de resultados e recursos desenvolvidos pela pesquisa em questão. Durante o período de estudos seis artigos foram publicados em simpósios, congressos e conferências internacionais e nacionais de grande importância em áreas citadas neste trabalho. Dois destes artigos estão relacionados diretamente ao desenvolvimento apresentado: PronounBP e AffectPT-br.

Resumidamente, as contribuições do presente no texto deste trabalho são:

- O desenvolvimento de um dicionário em português (LIWC_2015pt) para uso na versão 2015 do LIWC ;
- A avaliação empírica de resultados de comparações entre os dicionários existentes e o desenvolvido, consistindo em:

  - A realização uma comparação estatística dos resultados de análises de diferentes

versões de dicionários do LIWC;

- A avaliação de desempenho do LIWC_2015pt em tarefas de classificação utilizando MT.

Resultados

A pesquisa apresentada nesta dissertação indica melhora em relação à versão 2007 do LIWC em português, desenvolvida em 2011. Um total de 73 categorias são encontradas no LIWC_2015en. Conforme mencionado, é um recurso para processamento de textos que organiza palavras de acordo com os componentes emocionais, cognitivos e estruturais dos textos.

Para obter resultados de comparações, foram conduzidos experimentos tanto com o dicionário existente em inglês, quanto utilizando o LIWC_2007pt e o LIWC_2015pt. A avaliação empírica de resultados consistiu em realizar comparações de desempenho e observar os valores em diferentes tarefas de classificação. Também incluiu a comparação estatística dos resultados de análises de diferentes versões de dicionários do LIWC, conforme realizado em trabalhos relacionados.

Na maioria dos trabalhos relacionados encontrados, se observa que é realizada a comparação estatística dos resultados com textos bilíngues. A comparação com textos bilíngues em vários casos foi executada junto com experimentos da medida de desempenho em diferentes tarefas de classificação, como realizado neste trabalho. De forma semelhante, fizemos neste trabalho experimentos como nos estudos que apresentam diferentes traduções do LIWC_2015en e versões anteriores do dicionário.

Os experimentos de comparação estatística dos resultados com textos bilíngues foram realizados no sentido de estabelecer comparações entre os dicionários LIWC_2007pt e o LIWC_2015pt, que se trata de uma versão atualizada do dicionário compatível com a última versão do programa LIWC. Os resultados se mostraram satisfatórios nas comparações de desempenho pelos resultados de diferentes tarefas de classificações usando conjuntos com 190.000, 59.166 e 59.260 entradas de texto, assim como na comparação estatística das dimensões obtidas com as versões em português e inglês

para 2.823 textos bilíngues totalizando 150.000 sentenças alinhadas. Os experimentos indicaram que as palavras correspondem às diversas categorias do LIWC_2015en de forma apropriada e, sendo assim, permitem melhores resultados nas tarefas associadas às áreas de CA e AS.

Limitações do estudo

Ao longo dos estudos desenvolvidos, notamos uma grande dificuldade em obter conjuntos de dados para análise de sentimentos na língua portuguesa. Alguns dos conjuntos que encontramos contêm sequências de publicações repetidas e isso foi um fator que influenciou a decisão de não utilizá-los. Outro fator que impactaria negativamente os resultados com a abordagem que adotamos diz respeito aos conjuntos que disponibilizam conteúdo caracterizado pela rotulação somente quanto à presença de opinião sobre o item de interesse do estudo, como no caso do 'ReLi' ('REsenha de LIvros') (FREITAS et al., 2012). Isto ocorreria uma vez que a identificação de polaridade, neste conjunto, não estaria fazendo referência direta às palavras encontradas no texto publicado, se referindo apenas à opinião sobre o livro resenhado e sua polaridade.

Entendemos que em alguns casos, a disponibilização do conteúdo textual de redes sociais fere as regras de privacidade do serviço oferecido. Isso contribui para diminuir a quantidade de conjuntos que podem ser encontrados. A solução apresentada por (CAVALCANTE; MALHEIROS, 2017) se mostra como uma alternativa para criação de outros conjunto para análise de sentimentos e disponibilização de dados coletados de mensagens públicas na língua portuguesa presentes em redes sociais e pode ser sugerida para futuros pesquisadores.

Sobre o conjunto que usamos proveniente do MQD, observamos a vantagem dos textos poderem ser rotulados pelas emoções de Ekman, mas a desvantagem deste aspecto é que isto é realizado pelos próprios autores. Assim como o estado emocional, também não temos como garantir que outras informações do cadastro do usuário no sistema, como a idade, correspondem à realidade observável. Em outros casos, independente do rótulo ou de outras informações assinaladas pelos usuários, o conteúdo pode conter textos de outros autores, como fragmentos de obras literárias ou notícias. O

campo usado pelos autores para publicações pode ainda ser usado para guardar textos que são na verdade conversas com outras pessoas.

Quanto ao conjunto com o *corpus* paralelo, observamos como limitante o fato deste ser constituído de conteúdo de uma revista de notícias de conteúdo científico. Desta forma, entendemos que existam poucas palavras para serem detectadas e contadas em certas categorias, como as que analisam linguagem informal. Assim, seria interessante para futuras avaliações o uso de conjuntos bilíngues que contenham mais exemplos contemplando aspectos pessoais ou de uso de linguagem menos formal.

Quanto às palavras do dicionário, observamos que algumas palavras ainda comportam mais variações para distinção de sua função sintática, e.g., poderia ser incluída a forma adverbial 'visualmente', além dos adjetivos 'visual' e 'visuais'. Notamos também que alguns verbos ainda não possuem todas as conjugações possíveis. Esses são pontos que podem ser analisados para verificar se o aumento no número de palavras irá influenciar positivamente na obtenção de informações relevantes ou em resultados de testes.

Da mesma forma, identificamos que algumas categorias podem ser expandidas, considerando a quantidade de palavras atualmente incluída. A literatura encontrada sobre o LIWC e outros dicionários indica possibilidades utilizando expansão automática, via triangulação ou sequência a sequência. Entretanto, este também é um caso em que cabe a avaliação quanto o aumento do custo de processamento na comparação com resultados de desempenho nas tarefas a que o dicionário se destina.

Também será levado em consideração variações de palavras em relação ao vocabulário ortográfico para melhorar a identificação de palavras em redes sociais, transcrição de áudio e vídeo, além de conteúdo de aplicativos de troca instantânea de mensagens, considerando a informalidade observada nesses casos. Certas variações de palavras podem aparecer mais frequentemente quando o texto é produzido em um contexto de menor rigor em relação às normas ortográficas, como no caso de um registro mais pessoal, ou de publicação em redes sociais ou ainda em uma troca de mensagens rápida. Existe ainda a situação em que a rede social limita a quantidade de caracteres que podem ser utilizados para uma entrada, levando o usuário a optar por formas abreviadas. Quando o texto é submetido a uma ferramenta usada para AS, o resultado pode não ser tão bom por isso.

No presente trabalho incluímos pronomes pessoais com formas arcaicas e que diferem do vocabulário ortográfico (e.g., 'vc', 'vcs'), que já estavam no LIWC_2007pt, porém

em outras categorias não relacionadas a pronomes. Neste caso e em outros semelhantes, a palavra foi também incluída na categoria de Linguagem informal (‘informal’), conforme estrutura do LIWC_2015en. Antes de fazer a inclusão verificamos que havia equivalência com o que pode ser encontrado no LIWC_2015en.

Em trabalhos futuros, além de trabalharmos com as palavras atribuídas a categorias, pretendemos fazer uma comparação com métodos que usam conjuntos de recursos mais complexos. Também planejamos criar um algoritmo para processar N-gramas (ou seja, bigramas, trigramas, e assim por diante). Desta forma, buscamos obter do texto não apenas a informação de cada palavra isolada, mas também as informações relativas à estrutura do texto. Finalmente, avaliaremos o tempo de CPU usado por cada algoritmo.

Considerando também que, em um dos trabalhos publicados, nós propusemos um método para a análise psicolinguística de discursos orais extraídos de vídeos, também planejamos trabalhar com outros conteúdos além do conteúdo de textos. Pretendemos utilizar vídeos publicados em redes sociais e analisar o tom de voz e as expressões não verbais como uma abordagem complementar à analise textual. Os resultados podem ser usados para comparar as emoções expressas verbalmente (analisadas com nosso dicionário) àquelas que aparecem em expressões e gestos.

Bibliografia

AGARWAL, Apoorv; BIADSY, Fadi; MCKEOWN, Kathleen R. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In: PROCEEDINGS of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Atenas, Grécia: Association for Computational Linguistics, 2009. p. 24–32.

AIRES, Rachel et al. **Which classification algorithm works best with stylistic features of Portuguese in order to classify web texts according to users' needs?** São Carlos, Brasil, 2004.

ALTHOFF, Tim; JINDAL, Pranav; LESKOVEC, Jure. Online actions with offline impact: How online social networks influence online and offline user behavior. In: PROCEEDINGS of the Tenth ACM International Conference on Web Search and Data Mining (WSDM). Cambridge, Reino Unido: ACM, 2017. p. 537–546.

AZIZ, Wilker; SPECIA, Lucia. Fully Automatic Compilation of a Portuguese-English Parallel Corpus for Statistical Machine Translation. In: PROCEEDINGS of the 7th Brazilian Symposium in Information and Human Language Technology (STIL). Cuiabá, Brasil: Sociedade Brasileira de Computação, 2011.

BAKER, Mona. Corpora in translation studies: An overview and some suggestions for future research. **Target. International Journal of Translation Studies**, John Benjamins Publishing Company, v. 7, n. 2, p. 223–243, 1995.

BECKER, Karin; TUMITAN, Diego. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. In: SBBD - Simpósio Brasileiro de Banco de Dados. Recife, Brasil: Citeseer, 2013.

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para análise de sentimentos em mídias sociais. In: BRAZILIAN Symposium on Multimedia and the Web. Manaus, Brasil: ACM, 2015.

BERENGUER, Anabela et al. Are Smartphones Ubiquitous?: An in-depth survey of smartphone adoption by seniors. **IEEE Consumer Electronics Magazine**, IEEE, v. 6, n. 1, p. 104–110, 2017.

BJEKIĆ, Jovana et al. Psychometric evaluation of the Serbian dictionary for automatic text analysis-LIWCser. **Psihologija**, v. 47, n. 1, p. 5–32, 2014.

BOOT, Peter; ZIJLSTRA, Hanna; GEENEN, Rinie. The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. **Dutch Journal of Applied Linguistics**, John Benjamins Publishing Company, v. 6, n. 1, p. 65–76, 2017.

BOYD, Ryan L; PENNEBAKER, James W. A way with words: using language for psychological science in the modern era. **Consumer Psychology in a Social Media World**, p. 222–236, 2015.

BRADLEY, Margaret M; LANG, Peter J. **Affective norms for English words (ANEW): Instruction manual and affective ratings**. Gainesville, EUA, 1999.

BURK, Dan L. The Mereology of Digital Copyright. **Fordham Intellectual Property, Media & Entertainment Law Journal**, v. 18, p. 711–739, 2008.

CAETANO, Josemar Alves Caetano et al. Utilizando Análise de Sentimentos para Definição da Homofilia Política dos Usuários do Twitter durante a Eleição Presidencial Americana de 2016. In: VI Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017). São Paulo, Brasil: SBC, 2017.

CAMBRIA, Erik et al. New avenues in opinion mining and sentiment analysis. **IEEE Intelligent Systems**, IEEE, v. 28, n. 2, p. 15–21, 2013.

CARVALHO, Flavio; SANTOS, Gabriel dos; GUEDES, Gustavo Paiva. AffectPT-br: an Affective Lexicon based on LIWC 2015. In: 37TH International Conference of the Chilean Computer Science Society (SCCC 2018). Santiago, Chile: IEEE, 2018.

CARVALHO, Flavio et al. A dictionary of pronouns for Brazilian Portuguese. In: CONGRESSO Internacional de Informática Educativa (TISE). Brasília, Brasil: J. Sánchez, 2018.

CAVALCANTE, Paulo Emílio Costa; MALHEIROS, Yuri de Almeida. Um dataset para análise de sentimentos na língua portuguesa. Trabalho de Conclusão de Curso, Bacharel em Sistemas de Informação. Universidade Federal da Paraíba, 2017.

CHITICARIU, Laura; LI, Yunyao; REISS, Frederick R. Rule-based information extraction is dead! Long live rule-based information extraction systems! In: PROCEEDINGS of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Washington, EUA: Association for Computational Linguistics, 2013. p. 827–832.

COHEN, K Bretonnel; HUNTER, Lawrence. Getting started in text mining. **PLoS computational biology**, Public Library of Science, v. 4, n. 1, e20, 2008.

CROUX, Christophe; DEHON, Catherine. Influence functions of the Spearman and Kendall correlation measures. **Statistical methods & applications**, Springer, v. 19, n. 4, p. 497–515, 2010.

CROWL, Lawrence A. How to measure, present, and compare parallel performance. **IEEE Concurrency**, IEEE, n. 1, p. 9–25, 1994.

CRUZ, Pedro Parreira et al. Uma Revisão Sistemática sobre Léxicos Afetivos para o Português do Brasil. **Nuevas Ideas en Informática Educativa, TISE**, Fortaleza, Brasil, 2017.

D'AGOSTINO, Ralph B; BELANGER, Albert; D'AGOSTINO JR, Ralph B. A suggestion for using powerful and informative tests of normality. **The American Statistician**, Taylor & Francis, v. 44, n. 4, p. 316–321, 1990.

DE ALMEIDA, Napoleao Mendes. **Gramática metódica da língua portuguesa: curso único e completo**. São Paulo, Brasil: Saraiva, 1973.

DE CHOUDHURY, Munmun; COUNTS, Scott; HORVITZ, Eric. Social media as a measurement tool of depression in populations. In: PROCEEDINGS of the 5th Annual ACM Web Science Conference. Paris, França: ACM, 2013. p. 47–56.

EKMAN, Paul. An argument for basic emotions. **Cognition & emotion**, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992.

FAYYAD, Usama M et al. **Advances in knowledge discovery and data mining**. Cambridge, EUA: MIT Press, 1996. v. 21.

FAYYAD, Usama M; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic et al. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In: KDD. Portland, EUA: AAAI Press, 1996. v. 96, p. 82–88.

FELDMAN, Ronen; SANGER, James. **The text mining handbook: advanced approaches in analyzing unstructured data**. Cambridge, Reino Unido: Cambridge University Press, 2007.

FERSINI, Elisabetta; POZZI, Federico Alberto; MESSINA, Enza. Detecting irony and sarcasm in microblogs: The role of expressive signals and ensemble classifiers. In: DATA Science and Advanced Analytics (DSAA). Paris, França: IEEE, 2015. p. 1–8.

FILHO, Pedro P. Balage; PARDO, Thiago Alexandre Salgueiro; ALUÍSIO, Sandra 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. Fortaleza, Brasil: Sociedade Brasileira de Computação, 2013. p. 215–219.

FREITAS, Cláudia. Sobre a construção de um léxico da afetividade para o processamento computacional do português. **Revista Brasileira de Linguística Aplicada**, SciELO Brasil, v. 13, n. 4, 2013.

FREITAS, Cláudia et al. Vampiro que brilha... rá! Desafios na anotaçao de opinião em um corpus de resenhas de livros. **Encontro De Linguística de Corpus**, v. 11, p. 22, 2012.

GABRILOVICH, Evgeniy; MARKOVITCH, Shaul. Text categorization with many redundant features: using aggressive feature selection to make SVMs competitive with C4. 5. In: PROCEEDINGS of the twenty-first international conference on Machine learning. Banff, Canadá: ACM, 2004. p. 41.

GAO, Rui et al. Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In: INTERNATIONAL Conference on Brain and Health Informatics. Maebashi, Japão: Springer, 2013. p. 359–368.

GEHAN, Edmund A. A generalized Wilcoxon test for comparing arbitrarily singly-censored samples. **Biometrika**, Oxford University Press, v. 52, n. 1-2, p. 203–224, 1965.

GOLBECK, Jennifer; ROBLES, Cristina; TURNER, Karen. Predicting personality with social media. In: CHI'11 extended abstracts on human factors in computing systems. Vancouver, Canadá: ACM, 2011. p. 253–262.

GROSSMAN, Maura R; CORMACK, Gordon V. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. **Rich. JL & Tech.**, HeinOnline, v. 17, p. 1, 2010.

GUEDES, Gustavo Paiva et al. Gender Differences in the Use of Portuguese in Social Networks: Evidence from LIWC. In: PROCEEDINGS of the 22nd Brazilian Symposium on Multimedia and the Web. Teresina, Brasil: ACM, 2016. p. 339–342.

HALL, Mark et al. The WEKA data mining software: an update. **ACM SIGKDD explorations newsletter**, ACM, v. 11, n. 1, p. 10–18, 2009.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. EUA: Morgan Kaufmann Publishers, 2011.

HO, Tin Kam. Random decision forests. In: PROCEEDINGS of the third international conference on document analysis and recognition. Montreal, Canadá: IEEE, 1995. v. 1, p. 278–282.

HORNBY, Albert Sydney; WEHMEIER, Sally; ASHBY, Michael. Oxford advanced learner's dictionary. Oxford University Press Oxford, Oxford, Reino Unido, 2000.

HUANG, Chin-Lan et al. The development of the Chinese Linguistic Inquiry and Word Count dictionary. **Chinese Journal of Psychology**, Taiwanese Psychological Assn, Taiwan, 2012.

IKONOMAKIS, M; KOTSIANTIS, Sotiris; TAMPAKAS, V. Text classification using machine learning techniques. **WSEAS transactions on computers**, v. 4, n. 8, p. 966–974, 2005.

JELIER, Rob et al. Anni 2.0: a multipurpose text-mining tool for the life sciences. **Genome biology**, BioMed Central, v. 9, n. 6, r96, 2008.

KAILER, A; CHUNG, Cindy K. The Russian LIWC2007 dictionary. **LIWC.net**, Austin, TX, EUA, 2011.

KARP, Alan H; FLATT, Horace P. Measuring parallel processor performance. **Communications of the ACM**, Association for Computing Machinery, Inc., v. 33, n. 5, p. 539–544, 1990.

KENDALL, Maurice G. A new measure of rank correlation. **Biometrika**, JSTOR, v. 30, n. 1/2, p. 81–93, 1938.

KERBY, Dave S. The simple difference formula: An approach to teaching nonparametric correlation. **Comprehensive Psychology**, SAGE Publications Sage CA: Los Angeles, CA, v. 3, 11–it, 2014.

KLEINE-COSACK, Christian. Recognition and simulation of emotions. **Archived from the original on May**, v. 28, 2008.

KOHAVI, Ron et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: 2. IJCAI. Montreal, Canadá: AAAI Press, 1995. v. 14, p. 1137–1145.

LANDWEHR, Niels; HALL, Mark; FRANK, Eibe. Logistic model trees. **Machine learning**, Springer, v. 59, n. 1-2, p. 161–205, 2005.

LANGLEY, Pat; IBA, Wayne; THOMPSON, Kevin et al. An analysis of Bayesian classifiers. In: PROCEEDINGS of the tenth national conference on Artificial intelligence. San José, EUA: AAAI Press, 1992. v. 90, p. 223–228.

LEE, Chang H; SHIM, J; YOON, Aesun. The review about the development of Korean Linguistic Inquiry and Word Count. **Korean journal of cognitive science**, v. 16, n. 2, p. 93–121, 2005.

LIU, Bing. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan e Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

LIU, Haibin; KEŠELJ, Vlado. Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests. **Data & Knowledge Engineering**, Elsevier, v. 61, n. 2, p. 304–330, 2007.

LUFT, Celso Pedro et al. **Novo manual de português, gramática, ortografia oficial, literatura, redação, textos e testes**. São Paulo: Globo, 1997.

MANNING, Christopher D; SCHÜTZE, Hinrich. **Foundations of statistical natural language processing**. Cambridge, EUA: MIT press, 1999.

MASSÓ, Guillem et al. Generating New LIWC Dictionaries by Triangulation. In: ASIA Information Retrieval Symposium. Kuching, Malásia: Springer, 2013. p. 263–271.

MCCALLUM, Andrew; NIGAM, Kamal et al. A comparison of event models for naive bayes text classification. In: 1. AAAI-98 workshop on learning for text categorization. Madison, Wisconsin, EUA: Citeseer, 1998. v. 752, p. 41–48.

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MOWERY, Danielle; BRYAN, Craig; CONWAY, Mike. Feature studies to inform the classification of depressive symptoms from Twitter data for population health. In: PROCEEDINGS of the WSDM 2017 Workshop on Mining Online Health Reports. Cambridge, Reino Unido: ACM, 2017.

NAKAYAMA, Makoto; WAN, Yun. Is culture of origin associated with more expressions? An analysis of Yelp reviews on Japanese restaurants. **Tourism Management**, Elsevier, v. 66, p. 329–338, 2018.

NASCIMENTO, Gabriel; DUARTE, Fellipe; GUEDES, Gustavo Paiva. Handling Out-of-Vocabulary Words in Lexicons to Polarity Classification. In: PROCEEDINGS of the 17th Brazilian Symposium on Human Factors in Computing Systems. Belém, Brasil: ACM, 2018. p. 47.

NASSIRTOUSSI, Arman Khadjeh et al. Text mining for market prediction: A systematic review. **Expert Systems with Applications**, Elsevier, v. 41, n. 16, p. 7653–7670, 2014.

NOETHER, Gottfried E. Why Kendall Tau? **Teaching Statistics**, Wiley Online Library, v. 3, n. 2, p. 41–43, 1981.

OFEK, Nir et al. The Importance of Pronouns to Sentiment Analysis: Online Cancer Survivor Network Case Study. In: PROCEEDINGS of the 24th International Conference on World Wide Web. Florença, Itália: ACM, 2015. p. 83–84.

PANG, Bo; LEE, Lillian; VAITHYANATHAN, Shivakumar. Thumbs up?: sentiment classification using machine learning techniques. In: PROCEEDINGS of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Filadélfia, EUA: Association for Computational Linguistics, 2002. p. 79–86.

PENNEBAKER, James W. The secret life of pronouns. **New Scientist**, Elsevier, v. 211, n. 2828, p. 42–45, 2011.

_____. Using computer analyses to identify language style and aggressive intent: The secret life of function words. **Dynamics of Asymmetric Conflict**, Taylor & Francis, v. 4, n. 2, p. 92–102, 2011.

PENNEBAKER, James W; BOOTH, Roger J; FRANCIS, Martha E. Linguistic Inquiry and Word Count: LIWC [Computer software]. **Austin, TX: liwc. net**, 2007.

PENNEBAKER, James W; FRANCIS, Martha E; BOOTH, Roger J. Linguistic Inquiry and Word Count (LIWC): LIWC2001. **Mahway: Lawrence Erlbaum Associates**, v. 71, n. 2001, p. 2001, 2001.

PENNEBAKER, James W et al. **Linguistic Inquiry and Word Count: LIWC 2015 [Computer software]. Pennebaker Conglomerates**. Austin, TX, EUA: Inc, 2015.

PENNEBAKER, James W et al. **The development and psychometric properties of LIWC2015**. Austin, TX, EUA, 2015.

PETTIJOHN, Terry F; SACCO JR, Donald F. The language of lyrics: An analysis of popular Billboard songs across conditions of social and economic threat. **Journal of Language and Social Psychology**, Sage Publications Sage CA: Los Angeles, CA, v. 28, n. 3, p. 297–311, 2009.

PICARD, Rosalind Wright et al. Affective computing. Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995.

PICARD, Rosalind Wright. Affective computing: from laughter to IEEE. **IEEE Transactions on Affective Computing**, IEEE, v. 1, n. 1, p. 11–17, 2010.

PIOLAT, Annie et al. La version française du dictionnaire pour le LIWC: modalités de construction et exemples d'utilisation. **Psychologie française**, Elsevier, v. 56, n. 3, p. 145–159, 2011.

PRABOWO, Rudy; THELWALL, Mike. Sentiment analysis: A combined approach. **Journal of Informetrics**, Elsevier, v. 3, n. 2, p. 143–157, 2009.

QUINLAN, J Ross. **C4. 5: programs for machine learning**. San Mateo, CA, EUA: Elsevier Science, 2014.

_____. Induction of decision trees. **Machine learning**, Springer, v. 1, n. 1, p. 81–106, 1986.

RAMÍREZ-ESPARZA, Nairán et al. La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. **Revista mexicana de psicología**, Sociedad Mexicana de Psicología AC, v. 24, n. 1, 2007.

RAVI, Kumar; RAVI, Vadlamani. A novel automatic satire and irony detection using ensembled feature selection and data mining. **Knowledge-Based Systems**, Elsevier, v. 120, p. 15–33, 2017.

REIS, Julio CS et al. Uma abordagem multilíngue para análise de sentimentos. In: IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2015). Recife, Brasil: SBC, 2015.

RODRIGUES, Rafael Guimarães et al. Inferência de idade utilizando o LIWC: identificando potenciais predadores sexuais. In: VI Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017). São Paulo, Brasil: SBC, 2017.

RODRIGUES, Rafael Guimarães et al. TATMaster: Psycholinguistic Divergences in Automatically Translated Texts. In: PROCEEDINGS of the 23rd Brazillian Symposium on Multimedia and the Web. Canela, Brasil: ACM, 2017. p. 205–208.

RUDE, Stephanie; GORTNER, Eva-Maria; PENNEBAKER, James W. Language use of depressed and depression-vulnerable college students. **Cognition & Emotion**, Taylor & Francis, v. 18, n. 8, p. 1121–1133, 2004.

SAMANTA, Debasis; PANCHAL, Gaurang. Advances in Soft Computing. **Soft Computing Applications in Sensor Networks**, CRC Press, p. 21, 2016.

SATHYA, S; RAJENDRAN, N. A review on text mining techniques. **Int. J. Comput. Sci. Trends Technol**, v. 3, n. 5, p. 274–284, 2015.

SCHLER, Jonathan et al. Effects of age and gender on blogging. In: AAAI spring symposium: Computational approaches to analyzing weblogs. Palo Alto, EUA: AAAI, 2006. v. 6, p. 199–205.

SEBASTIANI, Fabrizio. Machine learning in automated text categorization. **ACM computing surveys (CSUR)**, ACM, v. 34, n. 1, p. 1–47, 2002.

SHAPIRO, Samuel Sanford; WILK, Martin B. An analysis of variance test for normality (complete samples). **Biometrika**, JSTOR, v. 52, n. 3/4, p. 591–611, 1965.

SHIBATA, Daisaku et al. Detecting Japanese Patients with Alzheimer's Disease based on Word Category Frequencies. In: PROCEEDINGS of the Clinical Natural Language Processing Workshop (ClinicalNLP). Osaka, Japan: The COLING 2016 Organizing Committee, 2016. p. 78–85.

SIMONS, Gary F; FENNIG, Charles D. Ethnologue: Languages of the world. **SIL, Dallas, Texas**, 2017.

STUDENT. The probable error of a mean. **Biometrika**, JSTOR, p. 1–25, 1908.

SVETNIK, Vladimir et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. **Journal of chemical information and computer sciences**, ACS Publications, v. 43, n. 6, p. 1947–1958, 2003.

TABOADA, Maite et al. Lexicon-based methods for sentiment analysis. **Computational linguistics**, MIT Press, v. 37, n. 2, p. 267–307, 2011.

TEMPLE, Mark. **Dicionário Oxford Escolar para estudantes brasileiros de inglês**. Oxford, Reino Unido: Oxford University Press, 2001.

TERRA, Ernani; NICOLA, José de. **Português: de olho no mundo do trabalho**. São Paulo: Scipione, 2004.

THODE, Henry C. **Testing for normality**. Bosa Roca, EUA: CRC press, 2002. v. 164.

TSENG, Yuen-Hsien; LIN, Chi-Jen; LIN, Yu-I. Text mining techniques for patent analysis. **Information Processing & Management**, Elsevier, v. 43, n. 5, p. 1216–1247, 2007.

TUMASJAN, Andranik et al. Predicting elections with Twitter: What 140 characters reveal about political sentiment. **Icwsm**, v. 10, n. 1, p. 178–185, 2010.

VALITUTTI, Alessandro; STRAPPARAVA, Carlo; STOCK, Oliviero. Developing affective lexical resources. **PsychNology Journal**, v. 2, n. 1, p. 61–83, 2004.

VAN RIJSBERGEN, Cornelis Joost. Information retrieval. Butterworth-Heinemann, 1979.

WANG, Sida; MANNING, Christopher D. Baselines and bigrams: Simple, good sentiment and topic classification. In: PROCEEDINGS of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers. Jeju, Coréia: Association for Computational Linguistics, 2012. v. 2, p. 90–94.

WISSEN, Leon van; BOOT, Peter. An Electronic Translation of the LIWC Dictionary into Dutch, 2017.

WOLF, Markus et al. Computergestützte quantitative textanalyse: äquivalenz und robustheit der deutschen version des Linguistic Inquiry and Word Count. **Diagnostica**, Hogrefe Verlag Göttingen, v. 54, n. 2, p. 85–98, 2008.

ZIJLSTRA, Hanna et al. De Nederlandse versie van de 'Linguistic Inquiry and Word Count'(LIWC). **Gedrag Gezond**, v. 32, p. 271–281, 2004.

Apêndice A

Tabela 14 – Comparação da quantidade de palavras do LIWC_2015en (2015), LIWC_2007pt (2007_pt) e LIWC_2015pt (2015_pt)

Categoria	Abreviação	Exemplos	2015	2007_pt	2015_pt
Dimensões					
Linguísticas					
Palavras de função	funct	Para, não, muito	491	5.512	1.426
Pronomes	pronoun	Eu, eles, próprio	153	128	234
Pronomes pessoais	ppron	Eu, conosco	93	54	79
1ª pessoa singular	i	Eu, comigo	24	7	9
1ª pessoa plural	we	Nós, nossa	12	8	11
2ª pessoa	you	Contigo, tu	30	25	36
3ª pessoa singular	shehe	Dela, ele, sua	17	16	17
3ª pessoa plural	they	Deles, elas, suas	11	11	15
Pron. impessoais	ipron	Nenhum, tanta	59	88	157
Artigos	article	A, um	3	10	29
Preposições	preps	Até, sem	74	69	73
Verbos auxiliares	auxverb	Estar, ter	141	1445	610
Advérbios Comuns	adverb	Aqui, nunca, só	140	139	526
Conjunções	conj	E, já, todavia	43	27	40
Gramática (outras)					
Negações	negate	Sequer, nada	62	21	36
Verbos	verb	Conquistar, ver	1.000	23.873	6.162
Adjetivos	adj	Incrível	764	NA	1.241
Comparações	compare	Antes, maior	317	NA	300
Interrogativas	interrog	Aonde, como	48	NA	23
Números	number	Mil, três	36	83	106

Categoria	Abreviação	Exemplos	2015	2007_pt	2015_pt
Quantificadores	quant	Alguns, pouco	77	622	160
Processos					
Psicológicos					
Proc. afetivos	affect	Admirável, agonia	1.393	28.475	2.105
Emoção positiva	posemo	Bem-estar, justo	620	12.878	863
Emoção negativa	negemo	Incômodo, solidão	744	15.115	1.213
Ansiedade	anx	Horror, medo	116	3.012	209
Raiva	anger	Irritada, mesquinho	230	6.867	315
Tristeza	sad	Lamentável, sofri	136	3.864	239
Social	social	Amiga, eles	756	13.634	1.445
Família	family	Filho, tia	118	96	93
Amigos	friend	Amada, chapa	95	679	61
Feminino	female	Mulher, neta	124	NA	152
Masculino	male	Mano, noivo	116	NA	125
Processos					
Cognitivos	cogproc	Porque, sabe	797	46.308	2.691
Discernimento	insight	Penso, sabe	259	18.683	1.095
Causal	cause	Efeito, porque	135	11.770	547
Discrepâncias	discrep	Devia, se	83	29.44	313
Tentativa	tentat	Confusa, quase	178	5.719	483
Certeza	certain	Nunca, sempre	113	3.428	425
Diferenciação	differ	Além, mas	81	NA	213
Processos					
Perceptivos	percept	Ouvir, ver, sentir	436	17.607	1.001
Ver	see	Aparente, imagem	126	4.634	451
Ouvir	hear	Barulhenta, soar	93	3.045	300
Sentir	feel	Cãibra, ternura	128	7.727	283

Categoria	Abreviação	Exemplos	2015	2007_pt	2015_pt
Processos					
Biológicos	bio	Comeu, dor	748	17.861	1.924
Corpo	body	Epitélio, sangue	215	4.766	271
Saúde	health	Chagas, muscular	294	7.003	1.296
Sexual	sexual	Acasalar, orgásmico	131	1.819	146
Ingerir	Ingest	Espaguete, laranja	184	11.805	336
Impulsos	drives	Ambiciosa, clã	1.103	NA	3.176
Afiliação	affiliation	Aliar, social	248	NA	601
Realização	achieve	Vencer, sucesso	213	9.865	896
Poder	power	Gerenciar, vitória	518	NA	1.146
Recompensa	reward	Bônus, vencer	120	NA	394
Risco	risk	Defesa, perigo	103	NA	511
Orientação					
Temporal	timeorient				
Foco passado	focuspast	Falou, passado	341	7.684	3.159
Foco presente	focuspresent	Agora, atual	424	4.715	1.416
Foco futuro	focusfuture	Amanhã, breve	97	268	1.456
Relatividade	relativ	Continuar, indo	974	24.966	3.220
Movimento	motion	Chegar, queda	325	13.641	1.989
Espaço	space	Acima, imenso	360	5.313	867
Tempo	time	Adiantado, hora	310	7.324	695
Preocupações					
Pessoais	persconc				
Trabalhos	work	Assessoria, baia	444	7.735	781
Lazer	leisure	Baile, pousada	296	6.331	556
Casa	home	Aluguel, quintal	100	2.019	145
Dinheiro	money	Ações, pensão	226	5.353	356
Religião	relig	Fé, santo, zen	174	2.066	343

Categoria	Abreviação	Exemplos	2015	2007_pt	2015_pt
Morte	death	Funeral, morreu	74	2.429	90
Linguagem					
Informal	informal	Buguei, td, vcs	380	NA	419
Palavrões	swear	Besta, cretino	131	14.041	174
'Netspeak'	netspeak	Bjs, hj, sqn	209	NA	212
Consentimento	assent	Boa, sim, show	36	58	40
Disfluências	nonflu	Ah, huh	19	14	38
Preenchimento	filler	Aí, né	14	12	17

Apêndice B

Tabela 15 – Resultados da análise do FAPESP-CORPUS para comparação entre os valores da Mediana, Mínimo (Mín) e Máximo (Máx) das porcentagens na totalidade das categorias do LIWC_2015en (En) e do LIWC_2015pt (Pt), e comparações entre LIWC_2015en x LIWC_2015pt (τ_1) e LIWC_2015en x LIWC_2007pt (τ_2), observando os coeficientes de correlação τ b de Kendall nos valores das categorias

Categoria		En			Pt			
(Abreviação)	Mediana	Mín	Máx	Mediana	Mín	Máx	τ_1	τ_2
D. linguíst.								
funct	46,90	9,52	57,75	46,81	10,74	58,25	**0,54**	0,46
pronoun	6,01	0,00	17,05	16,74	0,00	28,25	**0,53**	0,49
ppron	1,47	0,00	9,12	9,55	0,00	20,00	**0,26**	0,21
i	0,00	0,00	4,37	0,00	0,00	2,82	0,71	0,71
we	0,30	0,00	4,42	0,26	0,00	2,28	0,27	**0,33**
you	0,00	0,00	2,00	0,37	0,00	2,77	**0,16**	0,07
shehe	0,34	0,00	8,48	7,34	0,00	14,29	**0,24**	0,23
they	0,51	0,00	4,47	2,12	0,00	11,00	**0,37**	0,36
ipron	4,37	0,00	8,69	13,90	0,00	22,15	**0,52**	0,45
article	11,71	3,17	22,76	16,20	3,76	23,00	**0,39**	0,24
preps	17,52	3,97	23,78	23,25	8,72	30,00	**0,39**	0,35
auxverb	5,16	0,00	10,78	3,55	0,00	8,67	**0,55**	0,51
adverb	2,46	0,00	6,78	8,10	0,00	16,07	**0,53**	0,50
conj	4,76	0,00	11,02	8,16	0,00	16,82	0,38	**0,40**
negate	0,55	0,00	3,03	0,83	0,00	4,49	0,68	**0,72**
verb	9,00	0,00	17,23	7,92	0,00	17,16	0,46	**0,48**
adj	3,95	0,00	9,77	3,91	1,20	8,89	**0,47**	NA

Categoria		En			Pt			
(Abreviação)	Mediana	Mín	Máx	Mediana	Mín	Máx	τ_1	τ_2
compare	2,14	0,00	6,64	2,53	0,00	6,02	**0,56**	NA
interrog	1,19	0,00	5,69	3,79	0,00	11,00	**0,43**	NA
number	2,58	0,00	58,88	4,55	0,00	56,67	0,68	**0,73**
quant	1,74	0,00	5,47	2,01	0,00	6,76	**0,55**	0,33
Proc. Psic.								
affect	2,54	0,00	9,27	2,60	0,00	9,79	**0,59**	0,46
posemo	1,63	0,00	5,91	1,52	0,00	6,10	**0,56**	0,34
negemo	0,74	0,00	8,29	0,93	0,00	8,38	**0,67**	0,55
anx	0,12	0,00	4,20	0,09	0,00	3,87	**0,56**	0,42
anger	0,14	0,00	2,90	0,12	0,00	4,63	**0,50**	0,38
sad	0,14	0,00	2,65	0,20	0,00	2,11	**0,52**	0,38
social	4,26	0,00	15,34	3,93	0,00	13,06	**0,60**	0,37
family	0,00	0,00	4,57	0,08	0,00	5,76	0,54	**0,56**
friend	0,08	0,00	1,55	0,16	0,00	2,34	**0,27**	0,09
female	0,00	0,00	7,41	0,23	0,00	4,54	**0,38**	NA
male	0,30	0,00	8,48	0,48	0,00	4,62	**0,40**	NA
cogproc	8,21	0,93	17,49	10,09	0,47	21,62	**0,60**	0,44
insight	1,90	0,00	7,24	1,63	0,00	7,00	**0,59**	0,43
cause	2,19	0,00	6,67	3,20	0,00	7,33	0,42	**0,46**
discrep	0,69	0,00	3,96	1,70	0,00	5,17	**0,39**	0,31
tentat	1,38	0,00	5,00	1,89	0,00	4,85	0,48	0,48
certain	0,77	0,00	2,99	0,88	0,00	3,23	**0,52**	0,26
differ	1,81	0,00	4,79	2,79	0,00	8,11	**0,63**	NA
percept	1,24	0,00	11,02	1,83	0,00	8,85	0,40	**0,41**
see	0,49	0,00	6,84	0,95	0,00	7,69	0,40	**0,41**
hear	0,32	0,00	5,25	0,33	0,00	5,18	**0,49**	0,37
feel	0,14	0,00	3,38	0,47	0,00	4,42	**0,34**	0,27

Categoria		En			Pt			
(Abreviação)	Mediana	Mín	Máx	Mediana	Mín	Máx	τ_1	τ_2
bio	1,10	0,00	11,32	1,64	0,00	10,77	**0,64**	0,47
body	0,18	0,00	6,01	0,18	0,00	5,76	**0,61**	0,34
health	0,45	0,00	8,96	0,53	0,00	8,54	**0,64**	0,57
sexual	0,00	0,00	3,67	0,00	0,00	2,72	**0,58**	0,22
ingest	0,16	0,00	8,15	0,73	0,00	9,09	**0,39**	0,12
drives	5,66	0,00	14,84	7,09	0,00	15,76	**0,59**	NA
affiliation	1,15	0,00	6,95	1,07	0,00	5,97	**0,44**	NA
achieve	1,54	0,00	6,12	1,60	0,00	4,97	**0,56**	0,34
power	2,27	0,00	11,22	2,28	0,00	10,15	**0,54**	NA
reward	0,71	0,00	4,59	2,18	0,00	5,14	**0,33**	NA
risk	0,37	0,00	3,19	1,53	0,00	4,27	**0,31**	NA
Orient. Temp.								
focuspast	2,72	0,00	9,80	1,38	0,00	8,21	**0,63**	0,56
focuspresent	5,55	0,00	12,78	5,22	0,00	10,47	0,52	**0,54**
focusfuture	0,54	0,00	4,39	0,32	0,00	3,96	**0,48**	0,18
relativ	12,88	0,79	24,85	12,45	3,29	25,35	**0,57**	0,45
motion	1,34	0,00	4,28	3,11	0,00	8,38	**0,29**	0,27
space	7,82	0,79	19,03	7,34	2,00	18,69	**0,53**	0,42
time	3,49	0,00	11,96	4,84	0,00	16,99	**0,58**	0,33
Preocup. Pess.								
work	5,36	0,47	20,42	4,23	0,00	16,19	**0,77**	0,73
leisure	0,38	0,00	8,24	0,41	0,00	8,70	**0,44**	0,33
home	0,08	0,00	4,97	0,08	0,00	4,58	**0,49**	0,35
money	0,50	0,00	9,12	0,71	0,00	10,76	**0,60**	0,50
relig	0,03	0,00	7,24	0,06	0,00	6,32	**0,38**	0,36
death	0,00	0,00	3,24	0,00	0,00	3,26	**0,72**	0,41

Categoria		En			Pt			
(Abreviação)	Mediana	Mín	Máx	Mediana	Mín	Máx	τ_1	τ_2
informal	0,20	0,00	2,79	0,25	0,00	1,97	**0,01**	NA
swear	0,00	0,00	0,91	0,00	0,00	0,98	**0,10**	0,08
netspeak	0,09	0,00	2,79	0,00	0,00	1,14	**0,05**	NA
assent	0,00	0,00	1,00	0,09	0,00	1,38	**0,20**	0,09
nonflu	0,05	0,00	1,01	0,05	0,00	1,61	**0,07**	0,04
filler	0,00	0,00	0,34	0,10	0,00	1,61	0,06	**0,12**