Classificação de personalidade no Twitter: Uma análise sobre a viabilidade de transferência multicultural de aprendizagem

Arthur Pereira de Oliveira[1], **Marcos César da Rocha Seruffo**[1]

[1] Faculdade de Engenharia da Computação e Telecomunicações (FCT)
Instituto de Tecnologia – Universidade Federal do Pará (UFPA)
Caixa Postal 479 – 66075-110 – Belém – PA – Brasil

olvarthur@gmail.com, seruffo@ufpa.br

Abstract. *Extracting information about personality is a research object for studies with several applications, such as recommendation systems, recruitment processes, among others. This research field has well-established and efficient resources and methodologies, but with a high concentration of applications in data from the english language. This article contributes to the personality classification for languages that have a smaller number of resources, leveraging the use of techniques aimed at the english language, thus, allowing advances in personality classification technologies created for english-speaking audiences to reflect the advancement of other languages and cultures. For this, techniques of Data Mining, Natural Language Processing and Machine Learning with Word Embedding are used, exploring the correlation between personality traits and textual lexical properties, with data from Online Social Networks. The results obtained are satisfactory when compared with related research applied only in the English language, thus demonstrating the feasibility of wider use of techniques previously aimed at a single culture, indicating the possibility of overcoming the multicultural barrier present in the literature.*

Resumo. *Extrair informações sobre personalidade é objeto de estudo de pesquisas com diversas aplicações, como sistemas de recomendação, processos de recrutamento, entre outras. Esta área de pesquisa possui recursos e metodologias bem estabelecidas e eficientes, porém com alta concentração de aplicações em dados provenientes da língua inglesa. Este artigo contribui na classificação de personalidade para línguas que dispõem de um número menor de recursos, a partir da utilização de técnicas voltadas para a língua inglesa, assim, permitindo que os avanços de tecnologias de classificação de personalidade criados para o público falante do inglês reflitam no avanço de outras línguas e culturas. Para isto, são utilizadas técnicas de Mineração de Dados, Processamento de Linguagem Natural e Machine Learning com Word Embedding, explorando a correlação entre traços de personalidade e propriedades léxicas textuais, com dados oriundos de Redes Sociais Online. Os resultados obtidos são satisfatórios quando comparados com pesquisas correlatas aplicadas apenas na língua inglesa, assim, demonstrando a viabilidade de utilização mais ampla de técnicas antes voltadas a uma única cultura, indicando a possibilidade de superar a barreira multicultural presente na literatura.*

1. Introdução

Diversos serviços, como sistemas de recomendação, processos de recrutamento, propagandas direcionadas, design de softwares e outros oferecidos em nosso dia-a-dia, tem como diferencial a habilidade de customização e adaptação de uso de acordo com a individualidade de cada consumidor e cliente. Para isto, uma ferramente chave é a habilidade de inferir personalidade e suas nuances de acordo com as informações fornecidas de forma direta ou indireta.

A obtenção de classificação de personalidade, porém, quase sempre exige a aplicação de testes com a assistência de profissionais de psicologia, o que implica em alto custo financeiro e de recursos humanos. Avanços na área de Processamento de Linguagem Natural (PLN) e *Machine Learning*, tem buscado soluções para esta problemática no desenvolvimento de técnicas que possam inferir personalidade a partir da análise e processamento de dados textuais.

Estas técnicas tem fundamento na relação encontrada entre padrões lexicais e modelos de classificação de personalidade, como o Modelo dos Cinco Fatores (MCF) [Goldberg 1990], usado neste estudo, e objetivam fornecer informações sobre os autores dos textos de uma forma não intrusiva e com baixo custo humano e financeiro, sendo foco de diversos estudos, e.g. a tarefa compartilhada conduzida por Celli et al. [2013].

O modelo MCF, constituído por cinco domínios, nomeadamente, *Openness*, *Conscientiousness*, *Extroversion*, *Agreeableness* e *Neuroticism*, objetiva representar em escalas numéricas, características comportamentais de um indivíduo. Mais detalhadamente:

1. **Openness** – Pessoas com alto nível deste fator se interessam por novidades e variedades além de sensibilidade ao estado emocional próprio e de terceiros [Costa and Mccrae 1992];
2. **Conscientiousness** - Altos níveis deste fator está correlacionado com a necessidade por conquistas pessoais, bom gerenciamento de tempo e autocontrole [Costa and Mccrae 1992, Seidman 2013, John et al. 1999];
3. **Extroversion** - Um indivíduo com nível alto neste fator aprecia companhia de outros e estimula a interação social [Seidman 2013, Noguchi et al. 2006], além de apresentar gosto por festas e tendência a liderança em seus grupos sociais [Costa and Mccrae 1992]. Também são os indivíduos com maior propensão a uso de Redes Sociais Online [Seidman 2013];
4. **Agreeableness** - Está ligado a preferência a cooperatividade em detrimento da competição além de melhor aceitação a ideias e opiniões de terceiros [Costa and Mccrae 1992, Seidman 2013, Bayram and Aydemir 2017];
5. **Neuroticism** - Revela um nível moderadamente alto de emoções negativas, *stress* psicológico, maior insatisfação com aspectos pessoais, baixa auto-estima e insegurança [Costa and Mccrae 1992].

Com o crescimento de Redes Sociais Online (RSO) [Tankovska 2021], a possibilidade de extração e número de fontes de dados para estudo de classificação de personalidade se expande igualmente. Explorando este crescimento, torna-se possível a utilização de dados encontrados nestas plataformas para a potencialização de modelos de predição e classificação de personalidade. Esta abordagem corrobora com o objetivo de reduzir o

custo e tornar o processo menos invasivo, tendo em mente que dados de redes sociais, em sua maioria, estão disponíveis de forma pública através da Internet.

Com o desenvolvimento destas técnicas sendo um esforço global, os recursos e ferramentas disponíveis tendem a buscar o compartilhamento de ideias em uma linguagem que possa ser melhor compreendida mundialmente, resultando numa concentração de estudos e resultados feitos na língua inglesa [dos Santos et al. 2019], idioma que hoje possui um fator de globalização e padronização no que tange a comunicação entre pessoas de diferentes culturas.

Como resultado, linguagens diversas acabam por não usufruir da mesma gama de recursos para processamento de análises de dados na área de classificação e predição de personalidade [dos Santos et al. 2019]. Idealmente, a fim de alavancar estas linguagens e suas aplicações, o presente estudo busca formas de aproveitar ferramentas já criadas, objetivando sua aplicabilidade a estas linguagens, abrangendo além do português, alemão, mandarim, hindi e outras.

Portanto, este trabalho objetiva buscar soluções para a baixa disponibilidade de dados, informações, tecnologias e técnicas referentes a classificação de personalidade baseado em dados textuais na língua portuguesa do Brasil. Para isso, utilizou-se o *Twitter*, uma das RSO que apresenta maior crescimento e plataforma voltada exclusivamente ao compartilhamento de dados textuais de assuntos irrestritos [Newberry 2021], além de predominância da disponibilidade pública de dados, se demonstrando uma valiosa fonte de dados para a presente pesquisa, sendo, por estes motivos, a rede social escolhida como local de aplicação e avaliação da ferramenta proposta.

O trabalho está estruturado da seguinte forma: após a introdução à temática completa do trabalho feita na Seção 1, na Seção 2 são apresentados os estudos correlatos ao tema e de abordagem semelhante; na Seção 3 é apresentada a metodologia de estudo para as etapas de construção e aplicação do sistema proposto; na Seção 4 são apresentados os resultados extraídos do estudo de caso e discussão sobre estes; Na Seção 5 é feita a conclusão sobre a temática e debate sobre trabalhos futuros que possam endereçar as limitações encontradas durante a atual pesquisa.

2. Trabalhos Correlatos

O trabalho apresentado por Mairesse et al. [2007] é um dos primeiros da área e aborda o reconhecimento e classificação de personalidade em textos universitários e corpus de transcrição de discursos. O trabalho desenvolve um método que utiliza os conhecimentos psicolinguísticos abordados em *Linguistic Inquiry and Word Count* (LIWC) [Pennebaker et al. 2001] e o banco de dados psicolinguísticos do *Medical Research Council* (MRC) [Coltheart 1981] e compara o uso de técnicas de ranking, regressão e classificação.

Os resultados alcançados em [Hearst et al. 1998] demonstram uma precisão média variando entre 55% (em Realização) e 62% (em Abertura à experiência) utilizando *Support Vector Machine* (SVM) [Shevade et al. 1999]. Os autores destacam que a utilização de classificações feitas por especialistas, i.e. psicólogos analisando e classificando os dados, atingiram performance melhor em comparação a utilização de classificação a partir de auto-avaliação, método utilizado no presente estudo.

Em contraste com a utilização de modelos que se baseiam em características psicolinguísticas, e.g. LIWC, para extrair padrões, cita-se trabalhos que utilizam a extração de padrões direto do texto analisado. Esta metodologia necessita de uma grande quantidade de dados rotulados. Como exemplo, a proposta de Iacobelli et al. [2011] utilizando 2-grams unido a SVM para computar personalidades de escritores de blogs, atingindo uma performance melhor que LIWC devido a sua amplitude conceitual, destacam os autores.

Neste segmento, um número considerável de estudos tem utilizado redes sociais como fonte de dados para classificação de personalidade. O trabalho de tarefa compartilhada conduzido por Celli et al. [2013] aprofundou os esforços de classificação de personalidade ao disponibilizar datasets extraídos do *Facebook* e divulgou, entre outros trabalhos: a utilização de um espaço de padrões muito grande, incluindo informações sociais e demográficas; recursos lexicais; marcação *Part-of-Speech* unido aos modelos SVM e *Boosting* de *Machine Learning*; e utilização de 1-gram unido aos modelos SVM, Regressão Logística Bayesiana e Naïve Bayes Multinomial.

Mais recentemente, Arnoux et al. [2017] utilizaram *Word Embedding*, proposto por [Mikolov et al. 2013], técnica de representação de palavras através de vetores de valores contínuos, para extração de padrões e Processo Gaussiano (PG) [Rasmussen 2006] como modelo de classificação em textos publicados no *Twitter* (*tweets*), destacando que o modelo necessita de 8 vezes menos dados para alcançar a mesma performance do modelo estado-da-arte anterior, o qual utiliza LIWC [Schwartz et al. 2013]. Carducci et al. [2018] similarmente utilizou *Word Embedding*, porém unido a SVM, utilizando técnicas de transferência de aprendizagem ao usar como dados de treinamento, textos extraídos de publicações no *Facebook* e aplicando o modelo treinado para fazer predições em textos extraídos de publicações no *Twitter*.

Como explicitado por Santos, Ramos e Paraboni [2019], ainda que o estudo sobre classificação de personalidade em texto seja amplamente discutido e desenvolvida de forma contemporânea, existe uma quantidade consideravelmente reduzida de recursos disponíveis em línguas que não sejam a língua inglesa, como é o caso da língua portuguesa brasileira. Como efeito disto, se ressalta que, até o momento da escrita deste artigo, não há conhecimento dos autores sobre estudos que proponham simultaneamente validar transferência de aprendizagem entre plataformas e línguas (inglês e português).

O presente trabalho não visa contribuir com a melhoria de desempenho em relação ao estado-da-arte neste campo de estudo, mas com a utilização de técnicas voltadas para a língua inglesa em línguas com menos recursos[1], mais especificamente na língua portuguesa do Brasil, permitindo que o avanço de ferramentas em dados em inglês, possa resultar no avanço desta linguagem e outras que, igualmente, dispõem de menos recursos. Assim, para comparar os resultados, o presente estudo utilizou como referência as métricas apresentadas por trabalhos com metodologias e objetivos semelhantes, sendo os que foram propostos por Carducci et al. [2018] e Quercia et al. [2011].

3. Metodologia

Para o desenvolvimento da proposta de estudo, uma série de etapas foi planejada, visando permitir a reprodutibilidade do estudo. A Figura 1 apresenta as 5 etapas da metodologia,

[1]Entende-se como recursos, ferramentas, tais como o LIWC, que podem ser utilizadas para análises lexicais da língua, por exemplo.

bem como suas principais atividades e ferramentas utilizadas.

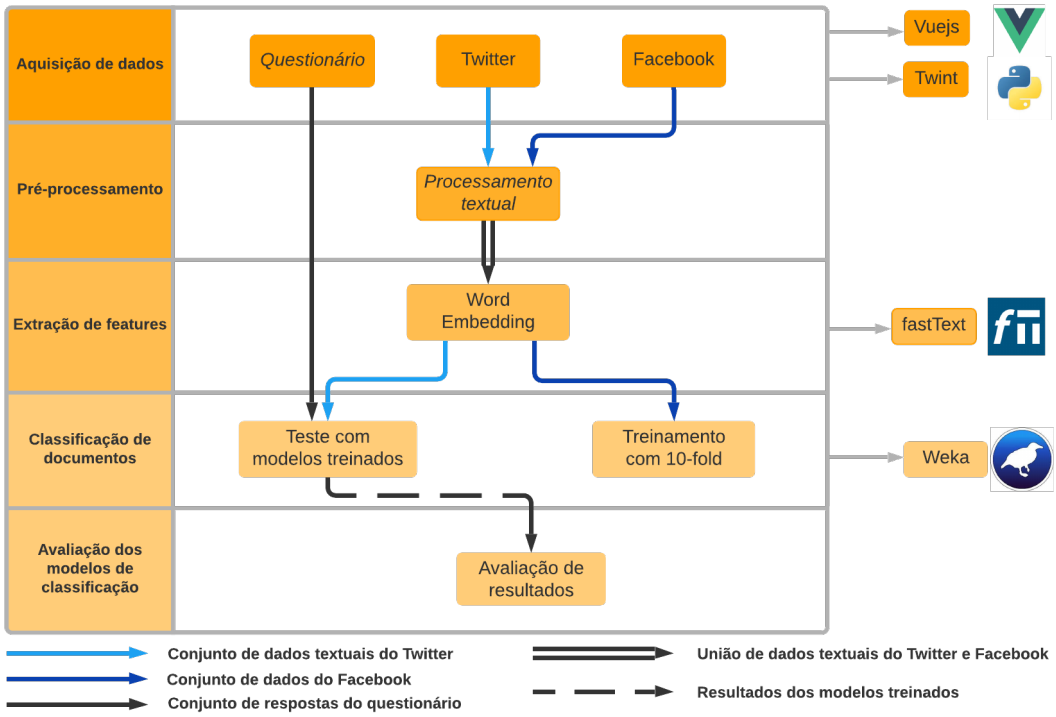

Figura 1. Etapas da metodologia aplicada

3.1. Aquisição de Dados

O conjunto de dados de treinamento é um subconjunto da amostra de MyPersonality[2], projeto criado por David Stillwell e Michal Kosinski, responsável pela aplicação de questionários de personalidade através de ferramentas do *Facebook*. O conjunto contém informações de 250 usuários e 9.917 publicações produzidas em forma de curtas sentenças sem restrição de assunto. Os participantes possuem dados textuais apenas na língua inglesa e suas pontuações na escala de fatores de personalidade foram obtidas através de autoavaliação.

Este conjunto foi disponibilizado inicialmente por Celli et al. [2013], também responsáveis pela anonimização e adição de classificações binárias derivadas da divisão mediana das pontuações. Baseado na análise apresentada por Carducci et al. [2018], se destacam as métricas apresentadas na Tabela 1.

[2]http://mypersonality.org

Tabela 1. Métricas analisadas na amostra MyPersonality

Métricas	Contagem
Usuários	250
Publicações	9917
Palavras	146.128
Palavras pós pré-processamento	72.896
Palavras únicas	15.470
Palavras únicas pós pré-processamento	15.185

O conjunto de dados de teste, por sua vez, foi criado durante o presente estudo e sua coleta se divide em dois momentos. Em um primeiro momento, através de um website criado com base no *template* disponibilizado por Enge et al. [2020], modificando o código fonte com o Framework Vuejs[3], de acordo com as necessidades da pesquisa. O website construído está disponivel em <https://tcc-delta.vercel.app/pt>. Com todas as explicações relacionadas a privacidade detalhadas no website, o participante forneceu seu nome de usuário usado no *Twitter* e as respostas para o questionário IPIP-NEO-120 [Johnson 2014], recebendo ao final as pontuações de personalidade dos 5 fatores e suas respectivas 6 facetas, além do detalhamento de cada faceta e suas implicações sobre o comportamento do participante.

Em um segundo momento, utilizando: o nome do usuário; o software *Twint*[4]; e um código desenvolvido em linguagem de programação *Python*[5], foi possível *crawlear* [6] a plataforma do *Twitter* a fim de obter os *tweets* públicos de cada participante. De forma análoga, os participantes aqui representados tem suas pontuações obtidas através de autoavaliação, porém possuem dados textuais apenas na língua portuguesa do Brasil. Ao fim da coleta foi realizada uma filtragem, para eliminar perfis privados, inexistentes ou sem *tweets*. Com isso, dos 20 perfis de usuários obtidos inicialmente, apenas 15 foram considerados resultando na Tabela 2.

Tabela 2. Métricas analisadas na amostra Twitter

Métricas	Contagem
Usuários	15
Publicações	9.228
Palavras	118.994
Palavras pós pré-processamento	68.191
Palavras únicas	23.863
Palavras únicas pós pré-processamento	15.903

3.2. Pré-processamento de Dados

Nesta etapa, os elementos textuais de ambos os conjuntos recebem as devidas modificações com o objetivo de eliminar o nível de ruído que possa vir a prejudicar a

[3]https://vuejs.org
[4]https://github.com/twintproject/twint
[5]https://www.python.org
[6]Coletar, de forma automatizada, conteúdo de uma ou mais páginas da web

análise. Os dados são primeiramente convertidos para letras minúsculas (*case folding*), em seguida são removidos as pontuações e *stopwords* [Lopes 2004], palavras que não adicionam valor semântico ao texto e é aplicada a tokenização, que é o processo de separação dos elementos que compõem o documento analisado.

Neste momento, ocorre uma separação entre os dados provenientes do *Facebook* (MyPersonality) e *Twitter*, neste último, é realizada: (i) remoção de url's; (ii) remoção de menções; (iii) remoção de *hashtags*, se mantendo a palavra utilizada já que esta contém informação importante sobre o conteúdo do *tweet*; (iv) e padronização, atividade que visa retornar os termos provenientes de abreviações e gírias para seus termos originais, permitindo sua análise correta pelo modelo.

3.3. Extração de Features

Nesta etapa, para os conjuntos de dados do *Facebook* e *Twitter*, se aplica a técnica de *Word Embedding*, escolhida pela capacidade de manter, em sua estrutura, informações semânticas e linguísticas de cada palavra [Mikolov et al. 2013].

Para implementação do *Word Embedding*, se utiliza o algoritmo *fastText* [Grave et al. 2018], escolhido por suportar nativamente as línguas inglesa e portuguesa, corroborando com a padronização de técnicas entre os conjuntos de dados utilizados. Em sua aplicação, as métricas de utilização são vetores de dimensão 300, com modelo 5-gram e janela de contexto igual a 5 (conjunto de métricas padrão). A fonte de dados para pré-treinamento deste algoritmo provem da plataforma *Common Crawl* [7].

3.4. Classificação de Documentos

Nesta etapa, com os dados textuais corretamente processados e relacionados com os dados de personalidade de cada participante, o conjunto de dados é separado novamente, de acordo com suas finalidades. Neste momento, o conjunto do *Facebook* é encaminhado para alimentar o modelo na fase de treinamento com validação *10-fold*, técnica utilizada para avaliar a capacidade de generalização do modelo [Refaeilzadeh et al. 2009]. Ao fim desta atividade, se obtém o modelo treinado, o qual recebe então os dados do conjunto *Twitter*, este, já unido com os dados referentes ao questionário aplicado.

Nestas atividades, os modelos de *Machine Learning* recebem os vetores resultantes do processo de *Word Embedding*. Com base nos estudos de: (i) Arnoux et al. [2017], que utiliza o PG; (ii) e Carducci et al. [2018] , que utiliza SVM, foram utilizados ambos modelos nesta proposta, aproveitando o levantamento da literatura e visando a comparação de resultados.

Embora PG e SVM tenham sido escolhidos como foco de análise no estudo, não se descartou a avaliação de performance de outros modelos. Assim, durante a avaliação de modelos de *Machine Learning*, se percebeu que o modelo de Regressão Linear (RL) [Kenney and Keeping 1962] apresentou resultados notoriamente satisfatórios quando comparado com os outros modelos supracitados, portanto, o modelo de RL foi incluído nos resultados, totalizando 3 modelos de classificação analisados e comparados no presente estudo.

[7] `https://commoncrawl.org/`

Além da linguagem de programação *Python*, a qual tem consideráveis recursos voltados para a prática de *Machine Learning* com destaque para as ferramentas scikit-learn [Pedregosa et al. 2011] e *Natural Language Toolkit* (NLTK) [Bird et al. 2009] utilizadas no estudo, se utilizou a ferramenta Weka [Witten et al. 2016], facilitando o uso dos modelos de classificação.

3.5. Avaliação dos Modelos de Classificação

Para avaliar os modelos de forma eficiente, se buscou referências que tenham apresentado resultados a partir de abordagem semelhante. Desta forma, os achados puderam ser comparados com Carducci et al. [2018] e Quercia et al. [2011].

A análise se baseou na grande semelhança entre as metodologias das pesquisas, porém, há a diferença desta proposta se preocupar com a adição da transferência de aprendizagem entre dados multiculturais. Este critério, portanto, corrobora com o objetivo apresentado ao fim da Seção 2, visando lançar luz sobre a possibilidade de utilização de técnicas voltadas para língua inglesa, em línguas com menor disponibilidade de recursos, a fim de permitir que o avanço de ferramentas restritas a uma cultura possa resultar no avanço de línguas como o português brasileiro e outras.

4. Resultados e Discussão

Esta Seção apresenta os resultados alcançados a partir dos modelos de *Machine Learning* utilizados, estabelecendo uma discussão sobre os objetivos propostos e os resultados obtidos no estudo.

4.1. Experimentação dos Modelos

Para que fosse possível encontrar as configurações otimizadas de performance na classificação, se buscou a exploração de diferentes algoritmos de *Machine Learning* e seus correspondentes valores de parâmetros que permitam maximizar suas performances. Com a finalidade de mensurar a qualidade dos resultados dos modelos, se utilizou o erro quadrático médio, parâmetro calculado utilizando a soma de todas as diferenças entre o valor alcançado em uma determinada classificação e o real valor desta classificação.

A escolha de tal métrica é motivada, além por ser aplicada em larga escala em estudos semelhantes, é apresentada por Carducci et al. [2018] , que expõe não apenas o seus resultados nesta métrica, mas também os resultados referentes ao trabalho de Quercia et al. [2011] . Para isso, se utilizou os achados relacionados a valores de parametrização e utilização de determinados modelos na literatura como base inicial, explorando diferentes configurações e padrões a partir desta base, através da experimentação.

O PG foi implementado através de código em Python com kernel RBF (*Radial basis function*) e alfa $1e{-}10$ através da biblioteca scikit-learn. A RL e SVM foram utilizados com o auxilio da ferramenta Weka. A RL foi implementada com o método de seleção de atributos M5 e ridge $1.0e^{-8}$. Em relação aos parâmetros de SVM, foram utilizados os mesmos descritos por Carducci et al. [2018].

Nesta etapa, para isolar os possíveis efeitos da aplicação de transferência de aprendizagem multicultural, os modelos foram alimentados apenas com a amostra de dados MyPersonality, com validação cruzada 10-fold.

Tabela 3. Valores de erro quadrático médio para amostra MyPersonality. OPN=*Openness*, CON=*Conscientiousness*, EXT=*Extroversion*, AGR=*Agreeableness*, NEU=*Neuroticism.*

Modelo	OPN	CON	EXT	AGR	NEU	MÉDIA
SVM	0,2164	0,3064	0,3170	0,3129	**0,2335**	0,2772
PG	0,4753	0,5802	0,5871	0,5856	0,5647	0,5585
RL	**0,1905**	**0,2465**	**0,2457**	**0,2514**	0,2345	**0,2337**

A Tabela 3 apresenta o valor de erro quadrático médio do MCF ao aplicar SVM, PG e RL. Nota-se a superioridade do modelo de RL, na maioria dos fatores, possuindo resultados 12% (OPN), 20% (CON e AGR) e 23% (EXT) melhores em comparação ao modelo SVM, segundo melhor colocado. A única exceção é o fator *Neuroticism* (NEU), o qual apresentou resultados aproximados para SVM e RL. Já o PG mostra resultados menos satisfatórios em comparação ao SVM e RL.

4.2. Avaliação dos Modelos

Com base na exploração de modelos de aprendizagem, foi avaliada a capacidade do poder de classificação da abordagem desta proposta, agora aplicando ao *dataset* criado com os dados do *Twitter* e seus usuários. Nesta abordagem, cada *tweet* é considerado como uma entrada independente de teste, tendo como vetor resultante a média de valores das palavras presentes na publicação após as atividades descritas na Seção 3.2.

Para diminuir um possível viés de determinada classe, no processo de *crawler* de *tweets*, se buscou balancear a base de dados mantendo uma quantidade similar de *tweets* recolhidos de cada usuário. Na Tabela 4, são comparados os resultados obtidos nesta proposta, para o MCF e a média dos fatores, com as outras referências utilizadas como base.

Tabela 4. Valores de erro quadrático médio para o MCF de acordo com datasets e estudos comparativos

Modelo	OPN	CON	EXT	AGR	NEU	MÉDIA
Modelo Proposto	0,4125	0,3711	0,3090	0,2441	0,2749	0,3223
[Quercia et al. 2011]	0,4761	0,5776	0,7744	0,6241	0,7225	0,6349
[Carducci et al. 2018]	0,3812	0,3129	0,3002	0,1319	0,2673	0,2787

Como demonstrado na Tabela 4, os resultados obtidos no modelo proposto alcançam uma média de 0,3223 entre os fatores do MCF, com *Agreeableness* atingindo o melhor resultado (0,2441). Destacando os resultados apresentados nos outros estudos, nota-se que o modelo proposto apresenta, em termos de média, uma melhora de 50% em relação a proposta de Quercia et al., [2011] entretanto, Carducci et al. [2018] aponta em seu estudo um modelo 15% melhor em relação a proposta deste artigo.

Ao analisar os resultados sobre a ótica do objetivo do estudo, i.e. buscar soluções para a baixa disponibilidade de dados, informações, tecnologias e técnicas referentes a classificação de personalidade baseada em dados textuais na língua portuguesa do Brasil, o desempenho do modelo proposto está em linha com o esperado, mantendo uma qualidade de classificação de personalidade compatível com os trabalhos citados. Assim, este

artigo contribui para a melhoria da análise da personalidade humana com transferência multicultural de aprendizagem, principal diferencial em relação a seus pares, os quais treinam e aplicam suas técnicas em dados provenientes de uma mesma cultura.

5. Conclusão

Este artigo apresenta a caracterização do indivíduo através do MCF, dando ênfase na demonstração desta personalidade com base em dados de RSO. É demonstrada a viabilidade de aplicação, entre conjuntos de dados multiculturais, da correlação entre conhecimentos lexicais extraídos de dados textuais de RSO e o modelo de personalidade. Esta viabilidade tem embasamento na semelhança entre os resultados encontrados na Tabela 3. Além disso, são comparados resultados oriundos da utilização de modelos de *Machine Learning* (SVM, PG e RL) nas bases de dados.

A performance do modelo proposto no presente estudo apresenta média de erro quadrático médio entre os 5 fatores de personalidade do MCF 15% maior que o apresentado por Carducci et al. [2018] . Considera-se que esta diferença, frente a maior complexidade trazida pela barreira multicultural ao presente trabalho, recai dentro de uma margem aceitável para o objetivo proposto, isto é, contribuir para línguas que dispõem de um número menor de recursos, a partir da utilização de técnicas voltadas para a língua inglesa, assim, permitindo que os avanços de tecnologias de classificação de personalidade criadas para o público falante do inglês, reflita no avanço de outras línguas e culturas.

Como limitação do trabalho é apontada a quantidade de participantes na coleta de dados realizada durante o estudo. Os autores apontam a dificuldade de encontrar pessoas para estudos desta natureza, principalmente motivado pelo momento de pandemia vivenciado, onde as pessoas estão menos acessíveis. Entretanto, por ser tratar de um modelo que poderá ser reproduzido em maior escala, a metodologia desenvolvida valida a proposta.

Como trabalhos futuros, um diferente processamento de postagens do *Twitter* pode ser aplicado. De forma mais especifica, pode ser utilizado como vetor resultante de *Word Embedding*, a concatenação entre os vetores de: (i) menor valor; (ii) maior valor;(iii) e média. Além disso, o trabalho pode ser expandido aumentando o número da participantes, assim, se obtendo uma maior diversidade de personalidades.

Referências

Arnoux, P.-H., Xu, A., Boyette, N., Mahmud, J., Akkiraju, R., and Sinha, V. (2017). 25 tweets to know you: A new model to predict personality with social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Bayram, N. and Aydemir, M. (2017). Decision-making styles and personality traits. 3.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media.

Carducci, G., Rizzo, G., Monti, D., Palumbo, E., and Morisio, M. (2018). Twitpersonality: Computing personality traits from tweets using word embeddings and supervised learning. *Information*, 9(5):127.

Celli, F., Pianesi, F., Stillwell, D., and Kosinski, M. (2013). Workshop on computational personality recognition: Shared task. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7.

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Costa, P. and Mccrae, R. (1992). Neo pi-r professional manual. *Psychological Assessment Resources*, 396.

dos Santos, W. R., Ramos, R. M., and Paraboni, I. (2019). Computational personality recognition from facebook text: psycholinguistic features, words and facets. *New Review of Hypermedia and Multimedia*, 25(4):268–287.

Enge, j. M., Gåsodden, G., Morten Amundsen, O., Sundt, P. A., and Moxnes, M. (2020). Big five personality test. `https://bigfive-test.com/`. Acessado em 01/12/2020.

Goldberg, L. R. (1990). An alternative "description of personality": the big-five factor structure. *Journal of personality and social psychology*, 59 6:1216–29.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., and Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28.

Iacobelli, F., Gill, A. J., Nowson, S., and Oberlander, J. (2011). Large scale personality classification of bloggers. In *international conference on affective computing and intelligent interaction*, pages 568–577. Springer.

John, O. P., Srivastava, S., et al. (1999). The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

Johnson, J. A. (2014). Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of Research in Personality*, 51:78–89.

Kenney, J. F. and Keeping, E. (1962). Linear regression and correlation. *Mathematics of statistics*, 1:252–285.

Lopes, M. C. S. (2004). Mineração de dados textuais utilizando técnicas de clustering para o idioma português. *Rio de Janeiro: sn*.

Mairesse, F., Walker, M. A., Mehl, M. R., and Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research*, 30:457–500.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Newberry, C. (2021). 36 twitter stats all marketers need to know in 2021. Acessado em 12/04/2021.

Noguchi, K., Gohm, C. L., and Dalsky, D. (2006). Cognitive tendencies of focusing on positive and negative information.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Quercia, D., Kosinski, M., Stillwell, D., and Crowcroft, J. (2011). Our twitter profiles, our selves: Predicting personality with twitter. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, pages 180–185. IEEE.

Rasmussen, C. E. (2006). Gaussian processes for machine learning. MIT Press.

Refaeilzadeh, P., Tang, L., and Liu, H. (2009). *Cross-Validation*, pages 532–538. Springer US, Boston, MA.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E., et al. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791.

Seidman, G. (2013). Self-presentation and belonging on facebook: How personality influences social media use and motivations. *Personality and individual differences*, 54(3):402–407.

Shevade, S., Keerthi, S., Bhattacharyya, C., and Murthy, K. (1999). Improvements to the smo algorithm for svm regression. In *IEEE Transactions on Neural Networks*.

Tankovska, H. (2021). Social media - statistics & facts. Acessado em 12/04/2021.

Witten, I. H., Frank, E., Hall, M., and Pal, C. (2016). The weka workbench. online appendix for "data mining: Practical machine learning tools and techniques". In *Morgan Kaufmann*. Fourth Edition, 2016.