UNISINOS

Programa de Pós-Graduação em

Computação Aplicada

Mestrado/Doutorado Acadêmico

Daniel Tamiosso

Personalidade e Redes Sociais: Agrupando e analisando características comportamentais de usuários de redes sociais a partir da combinação de traços de personalidade, dados demográficos e pegadas digitais

São Leopoldo, 2021

UNIVERSIDADE DO VALE DO RIO DOS SINOS — UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA
NÍVEL MESTRADO

DANIEL TAMIOSSO

PERSONALIDADE E REDES SOCIAIS: AGRUPANDO E ANALISANDO CARACTERÍSTICAS COMPORTAMENTAIS DE USUÁRIOS DE REDES SOCIAIS A PARTIR DA COMBINAÇÃO DE TRAÇOS DE PERSONALIDADE, DADOS DEMOGRÁFICOS E PEGADAS DIGITAIS

SÃO LEOPOLDO
2021

Daniel Tamiosso

PERSONALIDADE E REDES SOCIAIS: AGRUPANDO E ANALISANDO CARACTERÍSTICAS COMPORTAMENTAIS DE USUÁRIOS DE REDES SOCIAIS A PARTIR DA COMBINAÇÃO DE TRAÇOS DE PERSONALIDADE, DADOS DEMOGRÁFICOS E PEGADAS DIGITAIS

Dissertação apresentada como requisito parcial para a obtenção do título de Mestre pelo Programa de Pós-Graduação em Computação Aplicada da Universidade do Vale do Rio dos Sinos — UNISINOS

Orientador:
Prof. Dra. Patricia A. Jaques

São Leopoldo
2021

T158p Tamiosso, Daniel.
Personalidade e redes sociais : agrupando e analisando características comportamentais de usuários de redes sociais a partir da combinação de traços de personalidade, dados demográficos e pegadas digitais / Daniel Tamiosso. – 2021.
149 f. : il. ; 30 cm.

Dissertação (mestrado) – Universidade do Vale do Rio dos Sinos, Programa de Pós-Graduação em Computação Aplicada, 2021.
"Orientador: Prof. Dra. Patricia A. Jaques."

1. Redes sociais. 2. Computação da personalidade.
3. Racialização. 4. Modelo dos Cinco Grandes Fatores.
5. Clusterização. 6. Pegadas digitais. I. Título.
CDU 004

Dados Internacionais de Catalogação na Publicação (CIP)
(Bibliotecária: Amanda Schuster – CRB 10/2517)

Autorização Para Entrega Da Versão Final

Autorizo o/a aluno(a) **Daniel Tamiosso** a entregar a versão final da ***Dissertação*** sob o título de PERSONALIDADE E REDES SOCIAIS: AGRUPANDO E ANALISANDO CARACTERÍSTICAS COMPORTAMENTAIS DE USUÁRIOS DE REDES SOCIAIS A PARTIR DA COMBINAÇÃO DE TRAÇOS DE PERSONALIDADE, DADOS DEMOGRÁFICOS E PEGADAS DIGITAIS".

Saliento, ainda, que as correções sugeridas pela Banca foram atendidas.

São Leopoldo, 20 de Agosto de 2021.

Orientador(a): Prof(a). Dr (a) Patricia Augustin Jaques Maillard

Assinatura: .....

Aluno: Daniel Tamiosso

Assinatura:.....

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior Brasil (CAPES) - Código de Financiamento 001 / This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001

RESUMO

As redes sociais digitais estão se tornando cada vez mais populares e, com isso, elas oferecem uma plataforma massiva para a análise do comportamento humano em contextos mediados por computadores. O comportamento humano pode ser explorado pela análise do conjunto de rastros digitais criados pelas pessoas ao interagirem com as redes sociais. Esse rastro digital é definido como pegadas digitais. As pegadas digitais podem ser classificadas em ativas, quando produzidas de forma consentida, e passivas, quando produzidas de forma não intencional. É através delas que estudos podem explorar comportamento e interação social em larga escala. A descoberta de informações significativas e valiosas a partir de pegadas digitais deixadas nas redes sociais é realizada a partir das tecnologias de reconhecimento de padrões, bem como técnicas estatísticas e matemáticas; esta disciplina é referida como Mineração de Dados. Nesse contexto, este trabalho busca identificar perfis de usuários em redes sociais a partir do agrupamento de dados de comportamento em redes sociais (pegadas digitais), dados demográficos e informações socioafetivas (traços de personalidade), utilizando-se de técnicas de Mineração de Dados. Dessa forma, verifica-se a viabilidade na criação de grupos significativos considerando características socioafetivas e pegadas digitais, bem como disponibiliza-se uma análise qualitativa e quantitativa dos grupos produzidos, a fim de entender a qualidade dos grupos formados e a validade deles em relação aos conhecimentos revisados da Psicologia da Personalidade. Mais especificamente, são empregados algoritmos de aprendizado não supervisionados (clusterização), como K-Means e Spectral Clustering. Diferentemente de outros trabalhos na área de Computação da Personalidade que buscam identificar a personalidade dos usuários a partir das pegadas digitais usando algoritmos de aprendizado de máquina supervisionados, esse trabalho utiliza algoritmos de aprendizado de máquina não supervisionado, agrupando usuários de redes sociais com perfis semelhantes, a partir da coleta de pegadas digitais, dados demográficos e traços de personalidade. Isso permite entender as manifestações da personalidade de usuários de redes sociais pelo seu comportamento e características demográficas, ou seja, o papel que personalidades diferentes desempenham no comportamento dos usuários em redes sociais. Embora esse trabalho analise um grupo pequeno de usuários (157 participantes), pode-se verificar algumas correlações observadas na bibliografia relacionada, sendo um primeiro passo para propostas futuras a fim de trazer consciência sobre a relação das redes sociais, a Computação da Personalidade e os diversos campos subjacentes relacionados a dados estritamente pessoais e sensíveis. Esta pesquisa também traz como contribuição um novo conjunto de dados rotulados e com alta dimensionalidade (uma base de dados), os quais combinam dados comportamentais com características extraídas de pegadas digitais ativas e passivas, personalidade e informações demográficas de rede social na língua portuguesa.

Palavras-chave: Redes Sociais. Computação da Personalidade. Modelo dos Cinco Grandes Fatores. Clusterização. Pegadas Digitais.

ABSTRACT

Digital social networks are becoming more mainstream, offering a massive platform for analyzing human behavior in computer-mediated contexts. Algorithms can explore human behavior by analyzing digital footprints left by people when interacting with social networks. Digital footprints can be produced actively (in a consenting way) and passively (unintentionally). It is through them that studies can explore behavior and social interaction on a large scale. Thus, the discovery of essential and valuable information from digital footprints left on social networks is carried out using pattern recognition technologies and statistical and mathematical techniques; this discipline is referred to as data mining. This research seeks to identify user profiles in social networks by grouping behavior data in social networks (digital footprints), demographic data, and socio-affective profiles (personality traits). More specifically, unsupervised machine learning algorithms (clustering) such as K-means and Spectral Clustering are applied. Unlike other works on personality detection on social networks, the proposed work explores clustering techniques to group users with similar profiles by collecting their digital footprints, demographic data, and personality traits. From there, that work aims to understand the personality manifestations of social network users through their behavior, i.e, the role that different personalities play in the behavior of users on social networks. Although this work analyzes a small group of users (157 participants), some correlations observed in the related bibliography could be found. That work a first step for future incremental works in order to raise awareness about the relationship of social networks, Personality Computation and the several underlying fields related to strictly personal and sensitive data. This research also brings as a contribution a new set of labeled and high-dimensional data (a database), which combine behavioral data with characteristics extracted from active and passive digital footprints, personality and demographic information from a social network in Portuguese.

Keywords: Social Networks. Personality Computing. Big Five Personality Traits. Clustering. Digital Footrpints.

LISTA DE FIGURAS

1 Atualização de trabalhos pré-selecionados por área de interesse 34
2 Distribuição dos artigos selecionados por base 35
3 Recência dos trabalhos selecionados para análise 35
4 Número médio de curtidas de usuários caracterizados por diferentes níveis de abertura (traço de personalidade). O eixo y representa o intervalo interquartil médio do número de curtidas dos usuários no estudo de Kosinski et al. (2014). 41
5 Número médio de curtidas para usuários caracterizados por diferentes níveis de conscienciosidade. O eixo y representa o intervalo interquartil médio do número de curtidas dos usuários no estudo de Kosinski et al. (2014) 41
6 Dendrograma ilustrando a estrutura dos gostos musicais e a relação com o traço de personalidade da abertura. A estrutura foi produzida por Lambiotte e Kosinski (2014), usando um *cluster* hierárquico com as curtidas mais populares do Facebook na categoria musical. A escala de cores representa o traço de personalidade de abertura à experiência média, variando de conservador (azul ciano) para liberal (magenta) 43
7 Modelo de classificação das motivações de deixar pegadas digitais desenvolvido por Muhammad, Dey e Weerakkody (2018) 44
8 Na esquerda observa-se fotos de perfis de pessoas com fatores altos de extroversão, ao contrário do conjunto de fotos da direita que possuem baixa extroversão. As imagens foram disponibilizadas no estudo de Segalin et al. (2017) e o embaçamento foi utilizado por questões de privacidade. 45
9 Investimento em publicidade digital no Facebook (STATISTA, ????) 56
10 Exemplos de anúncios destinados a públicos caracterizados por alto e baixo índice de extroversão 58
11 Exemplos de anúncios destinados a públicos caracterizados por alto e baixo índice de abertura à experiência 58
12 Como funciona a segmentação psicológica? Figura projetada a partir da pesquisa de Matz, Appel e Kosinski (2020) 59

13 As duas fases do Trabalho Proposto 66
14 Wedy e sua linha do tempo social 69
15 Wedy Marketplace - uma rede de transações financeiras com interações sociais 70
16 Volume de publicações na linha do tempo da Wedy 71
17 Volume de produtos anunciados na linha do tempo da Wedy 72
18 Volume de curtidas nas publicações na linha do tempo da Wedy 72
19 Volume de comentários nas publicações na linha do tempo da Wedy 72
20 Volume de visualizações únicas nas publicações na linha do tempo da Wedy 72
21 Queda acelerada (+90%) no volume de dados disponibilizados pela rede social Wedy durante a pandemia da Covid-19 73
22 Exemplo de resultado apresentado ao usuário após envio das respostas do questionário. Esses dados poderão ser adaptados ao contexto inerente à rede social. 81
23 Exemplo de etapa de coleta de dados de personalidade a partir de formulário online. 82
24 Exemplo de uma publicação na rede social da Wedy, quando acessada por um navegador de Internet de um computador. 82

25 Estatísticas descritivas do conjunto de dados comportamentais dos usuários da rede social Wedy. 83

26 Correlação do conjunto de dados comportamental 86
27 Correlação do conjunto de dados demográficos 87
28 Correlação do conjunto de dados de personalidade 88
29 Correlação do conjunto de dados unificado: comportamento, demografia e personalidade 89
30 Uma hipotética relação entre um usuário e os seus conteúdos preferidos na rede social Wedy. 90
31 Principais palavras publicadas pelos usuários 90
32 Tópicos de maior interesse coletados a partir de pegadas digitais passivas . . 91
33 Tópicos de menor interesse coletados a partir de pegadas digitais passivas . . 91
34 Distribuição por gênero dos usuários da rede social com traços de personalidade coletados 92
35 Densidade de usuários por região brasileira com traços de personalidade coletados 93
36 Características dos eventos organizados pelos usuários com traços de personalidade coletados 94
37 Exemplo de diferença no traços de personalidade de "abertura à experiência"entre os usuários por gênero 95
38 Histograma dos traços de personalidade no conjunto de dados combinado . . 96
39 Distribuição comportamental nas principais atividades da rede social no conjunto de dados combinado 96
40 Novas dimensões e dimensões transformadas após a aplicação de diferentes técnicas de discretização 103
41 Resultado da aplicação de escalonamento para cada caso de assimetria transformando a distribuição de cada uma delas em normal. 105
42 Curva de limite para escolha de características com maior variância entre o conjunto de dados 108
43 Os valores DBI e SC para as várias formações de agrupamento do conjunto de dados usando K-means, K-medoides, Agglomerative Clustering e Spectral Clustering 109
44 Coeficiente de importância de cada dimensão na formação dos agrupamentos para as Principais Dimensões 112
45 Coeficiente de importância dos traços de personalidade na formação dos agrupamentos para as Principais Dimensões 112
46 Análise de distribuição dos clusters no conjunto Principais Dimensões: Traços de Personalidade 113
47 Análise de distribuição dos clusters no conjunto Principais Dimensões: Pegadas Digitais Comportamentais 115
48 Análise de distribuição dos clusters no conjunto Principais Dimensões: Demografia 115
49 Análise de distribuição dos clusters no conjunto Principais Dimensões: Análise Linguística 116
50 Análise de distribuição dos clusters no conjunto Principais Dimensões: Pegadas Digitais Passivas de Comportamento (categoria de conteúdo) 117
51 Distribuição dos traços de personalidade no modelo dos Cinco Grande Fatores para o Cluster 1 118

52 Distribuição dos traços de personalidade no modelo dos Cinco Grande Fatores para o Cluster 2 118
53 Traços de Personalidade Índices de Comportamento: Distribuição dos agrupamentos 119

LISTA DE TABELAS

1 Trabalhos relacionados selecionados 36
2 Tabela comparativa de dimensões utilizadas em trabalhos relacionados ao estudo de personalidade em redes sociais 50
3 Uma visão geral das características dos conjuntos de dados 78
4 Análise descritiva do conjunto de dados comportamentais 85
5 Análise descritiva do conjunto de dados demográficos 85
6 Análise descritiva do conjunto de dados personalidade 85
7 Dimensões com maiores desvios padrões 97
8 Dimensões com maiores desvios padrões após remoção de *outliers* 98
9 Dimensões com ausência de valores 99
10 Traços de personalidade agrupados em quartis por fator de intensidade no contexto do conjunto de dados 102
11 Comparação dos melhores índices em diferentes técnicas de clusterização para os três conjuntos de características 110
12 Representação sintética das características de cada Cluster. 114
13 Representação sintética das características de cada Cluster. 116
14 Cronograma de atividades da dissertação 139
15 Listagem de todas colunas presentes no conjunto de dados 140

LISTA DE SIGLAS

IA	Inteligência Artificial
CA	Computação Afetiva
GPS	*Global Positioning System*
DBI	*Davies–Bouldin Index*
SC	*Silhouette Coefficient*
TF-IDF	*Term Frequency–Inverse Document Frequency*
LIWC	Linguistic Inquiry and Word Count
MRC	Machine Reading Comprehension
POS	Part-of-Speech Tagging
NEO	Neuroticism-Extraversion-Openness
IGFP-5	Inventário dos 5 Grandes Fatores de Personalidade
LGPD	Lei Geral de Proteção de Dados
GDPR	*General Data Protection Regulation*
MAPE	Mean absolute percentage error
F1	Fase 1 da pesquisa
F2	Fase 2 da pesquisa
ER5FP	Escala Reduzida de Cinco Grandes Fatores de Personalidade

SUMÁRIO

1 INTRODUÇÃO **13**
1.1 Objetivos 15
1.2 Organização do documento 16

2 AFETO E COMPUTAÇÃO **17**
2.1 Computação Afetiva 17
2.2 Personalidade 20
2.3 Modelo dos Cinco Grande Fatores 21

3 PEGADAS DIGITAIS **24**
3.1 Pegadas Digitais Ativas 25
3.2 Pegadas Digitais Passivas 27

4 MINERAÇÃO DE DADOS E ALGORITMOS DE CLUSTERIZAÇÃO **28**
4.1 Algoritmos de Clusterização Selecionados 30

5 TRABALHOS RELACIONADOS **33**
5.1 Resultados 38
5.1.1 (QP_1) Traços de personalidade e comportamentos de usuário em redes sociais 38
5.1.2 (QP_2) A relação entre pegadas digitais em redes sociais e a personalidade de usuários 42
5.1.3 (QP_3) Métodos e técnicas de mineração de dados e aprendizado de máquina empregadas para detectar automaticamente a personalidade 45
5.1.4 (QP_4) Métodos e técnicas de mineração de dados e aprendizado de máquina empregadas para agrupar usuários de redes sociais de acordo com seu comportamento e personalidade 47
5.1.5 (QP_5) Dimensionalidade de dados utilizados nas pesquisas de agrupamento e inferência de personalidade de usuários em redes sociais 48
5.1.6 (QP_6) A detecção da personalidade como ferramenta para influenciar o comportamento de usuários nas redes sociais 55
5.1.7 (QS_1) Questões modernas relacionadas à privacidade no uso de pegadas digitais ativas e passivas para objetivos comerciais e de pesquisa 57
5.1.8 Discussão dos Resultados 61

6 TRABALHO DESENVOLVIDO **64**
6.1 Etapas do Trabalho 64
6.2 Viabilidade da Proposta 66
6.3 Materiais 67
6.3.1 Instrumentos psicológicos 67
6.3.2 A rede social Wedy 68
6.4 Conjunto de Dados 69
6.4.1 Limitações no conjunto de dados 71
6.4.2 Dados de personalidade 73
6.4.3 Dados comportamentais 74
6.4.4 Dados demográficos 76
6.5 Método 77

6.5.1 Coleta de dados . . . 77
6.5.2 Pré-processamento dos dados . . . 77
6.5.3 Clusterização aplicada . . . 78
6.5.4 Análise descritiva . . . 79
6.5.5 Questões de privacidade . . . 80

7 ANÁLISE E RESULTADOS . . . 84
7.1 Processo de Clusterização . . . 84
7.2 Coleta de dados . . . 84
7.3 Pré-processamento de dados . . . 92
7.3.1 Limpeza dos dados . . . 97
7.3.2 Imputação de valores ausentes . . . 97
7.3.3 Discretização . . . 99
7.3.4 Escalonamento dos dados . . . 104
7.4 Clusterização aplicada . . . 106
7.4.1 Algoritmos selecionados . . . 106
7.4.2 Seleção de características exploradas . . . 107
7.4.3 Número de agrupamentos validados . . . 107
7.4.4 Análise descritiva dos agrupamentos gerados . . . 108
7.5 Discussão . . . 120

8 CONCLUSÕES . . . 123
8.1 Ameaças à Validade dos Resultados . . . 124
8.2 Trabalhos Futuros . . . 125

REFERÊNCIAS . . . 127

ANEXO A MATERIAIS UTILIZADOS NA PESQUISA . . . 136
A.1 Questionário dos Cinco Grande Fatores . . . 136

ANEXO B CRONOGRAMA . . . 138
B.1 Principais etapas da dissertação . . . 138

ANEXO C DADOS . . . 140
C.1 Síntese do conjunto de dados unificado da pesquisa com todas as dimensões . . . 140
C.2 Distribuição de valores de cada dimensão do conjunto Principais Dimensões em seus dois agrupamentos . . . 145
C.3 Distribuição de valores de cada dimensão do conjunto Traços de Personalidade & Principais índices de comportamento em seus dois agrupamentos . . . 149

]

1 INTRODUÇÃO

As redes sociais digitais estão se tornando cada vez mais populares: 3,02 bilhões de pessoas estarão ativas em mídia social até 2021 (STATISTA, 2020). Esse aumento sem precedentes oferece uma plataforma massiva para a análise do comportamento humano em contextos mediados por computadores (GAVRILOVA, 2018). Isso se deve ao alto crescimento de dispositivos habilitados para Internet que, combinado ao comportamento humano completamente conectado, aumenta a proliferação de dados, deixando registros chamados de pegadas digitais. Dessa forma, as pegadas digitais são o conjunto de rastros digitais criados por pessoas ao interagirem com os canais ou dispositivos digitais, tais como curtidas no Facebook, compartilhamento de mensagens, etc. As pegadas digitais podem ser classificadas em ativas e passivas. Quando ativas, elas são produzidas com o consentimento dos usuários, como o ato de publicar um conteúdo público em uma rede social, o preenchimento de formulários e outras informações que o usuário publica intencionalmente. As pegadas digitais passivas referem-se aos dados deixados de forma não intencional, como o rastro de navegação entre conteúdos de uma rede social, o tempo da sessão, a frequência de uso e outras informações não publicas. Esses registros são potencialmente utilizados para prever traços humanos íntimos, como perfis de personalidade (LAMBIOTTE; KOSINSKI, 2014).

O grande conjunto de atividades digitais, também chamado de “Grandes Dados Sociais”, está fortalecendo a pesquisa científica, criando uma transição de estudos de pequena escala (geralmente empregam questionários de autorrelato ou observações e experimentos baseados em laboratório), para estudos remotos em grande escala (LAMBIOTTE; KOSINSKI, 2014). Trabalhos científicos têm mostrado o progresso da *Personality Computing* – um campo de pesquisa relacionado à Inteligência Artificial e à Psicologia da Personalidade – que estuda a personalidade por meio de técnicas computacionais. Alguns experimentos demonstram que as máquinas podem reconhecer tão bem a personalidade quanto os humanos ao analisar as pegadas digitais sociais de usuários, como as curtidas em páginas do Facebook (YOUYOU; KOSINSKI; STILLWELL, 2015).

Essa também é uma oportunidade para muitos setores comerciais, da publicidade, assim como para o desenvolvimento de produtos digitais (CHEN; PAVLOV; CANNY, 2009). Além disso, campanhas políticas estão transformando as redes sociais em um palco eleitoral, com muitas preocupações relacionadas à privacidade dos seus usuários (GRANVILLE, 2018), devido à utilização da inferência de traços de personalidade das pessoas como tentativa de manipulá-las e influenciá-las (YOUYOU; KOSINSKI; STILLWELL, 2015). Os negócios em geral estão identificando estruturas e padrões para entender e criar estratégias de influência em redes sociais. Eles estão usando o modelo de personalidade dos Cinco Grandes Fatores – um modelo reconhecido para avaliar os traços de personalidade (KUSS; GRIFFITHS, 2011) – para fornecer mensagens de marketing altamente personalizadas e ajustar seus produtos para melhor adequação ao perfil psicológico de cada usuário (YOUYOU; KOSINSKI; STILLWELL, 2015).

Ou seja, um dos principais *insights* oferecidos pelas pegadas digitais ativas e passivas em redes sociais refere-se à previsibilidade dos traços psicológicos individuais.

A descoberta de informações significativas e valiosas sobre essas pegadas digitais, especialmente deixadas nas redes sociais, é realizada a partir de tecnologias de reconhecimento de padrões, bem como técnicas estatísticas e matemáticas. Esta disciplina é referida como Mineração de Dados (HAND, 2006). De forma geral, existem muitas oportunidades para algoritmos de alta eficácia no julgamento da personalidade humana (YOUYOU; KOSINSKI; STILLWELL, 2015). Nesse contexto, os trabalhos que visam a detecção de traços de personalidade utilizam essencialmente a abordagem supervisionada. Entretanto, como a pesquisa desenvolvida nessa dissertação é direcionada para a descoberta de padrões comportamentais, demográficos e traços de personalidade previamente coletados, a escolha é pela utilização de algoritmos de aprendizado não supervisionado, e especificamente algoritmos de clusterização. Clusterização é a tarefa de dividir uma população ou pontos de dados em vários grupos, de modo que os pontos de dados nos mesmos grupos sejam mais semelhantes a outros pontos de dados no mesmo grupo do que os de outros grupos. Ou seja, como essa pesquisa estuda como agrupar dados de personalidade, comportamento e demografia sem conhecer os rótulos de cada potencial agrupamento, o estudo de algoritmos não supervisionados é fundamental para essa pesquisa.

A maioria dos pesquisadores realiza seus estudos com base no Facebook. No entanto, existem questões de privacidade sobre as informações provenientes das redes sociais (KOSINSKI M.; HANCOCK, 2005), o que tem gerado uma limitação de estudo sobre a personalidade dos usuários a partir da mineração de dados nas redes sociais. Políticas de privacidade cada vez mais rígidas impedem o acesso a esses dados em grande volume e restringem a coleta apenas ao conteúdo público consentido por seus usuários. Dessa maneira, os pesquisadores estão fazendo seus estudos com conjuntos de dados desatualizados ou com amostragens pequenas e limitadas a poucas dimensões.

Esta dissertação explora uma alternativa de estudo sobre as relações entre o comportamento humano em redes sociais e seus traços de personalidade, sob uma ótica que contempla a coleta e a manipulação de pegadas digitais em uma rede social de língua portuguesa, com os objetivos de (1) desenvolver agrupamentos, a partir de técnicas de clusterização de Mineração de Dados, considerando comportamento, personalidade e dados demográficos, permitindo a verificação da possibilidade de criação de grupos significativos considerando características socioafetivas e pegadas digitais passivas, bem como a (2) a análise qualitativa e quantitativa dos grupos produzidos, a fim de entender a qualidade dos grupos formados e a validade deles em relação aos conhecimentos revisados da Psicologia da Personalidade.

Para a criação de agrupamentos e a avaliação deles, esse trabalho utiliza-se de técnicas de clusterização, como K-Means e Spectral Clustering. Além da coleta, os passos metodológicos envolvem explorar os dados disponíveis, realizar o pré-processamento deles, modelar algoritmos de clusterização e avaliar os grupos resultantes com métricas específicas. Dessa forma, seguindo essa metodologia, é possível avaliar quais os agrupamentos desenvolvidos e detalha-

dos nas seções sequentes, possuem melhor validade estatística.

O trabalho também realiza uma análise descritiva e qualitativa dos agrupamentos produzidos, embora em um conjunto de dados pequeno e com alta dimensionalidade, a fim de entender melhor como a personalidade reflete-se no comportamento e nas características demográficas das amostras de usuários estudados. Dessa forma, pode-se explorar e entender se dados de personalidade poderiam ser agrupados de forma valiosa quando colocados de forma igualitária a dados comportamentais e demográficos extraídos de pegadas digitais ativas e, principalmente, passivas, e como esses se relacionam.

Como diferencial de modelagem em relação aos demais trabalhos relacionados, está o acesso a um conjunto de dados que não restringe-se apenas a pegadas digitais ativas, mas também as pegadas digitais passivas. Como complementar aos dados disponibilizados pelo produto digital Wedy [1], que possui uma rede social especializada no planejamento de eventos de casamento, foi realizada a coleta de dados socioafetivos, especificamente traços de personalidade do modelo dos Cinco Grandes Fatores, utilizando-se de um questionário de escala reduzida de auto relato (ER5FP), seguindo as políticas de privacidade recomendadas e aceitas pelos usuários do produto.

Esta pesquisa também introduz um novo conjunto de dados rotulados com alta dimensionalidade referentes à combinação de dados comportamentais com características extraídas das pegadas digitais ativas e passivas, personalidade e informações demográficas de rede social na língua portuguesa. Esse conjunto de dados, com aproximadamente 450 dimensões e, ao final dessa pesquisa, com 157 participantes selecionados, potencialmente permite que novas contribuições sejam feitas para avançar o estado da arte no estudo da Computação da Personalidade, aprofundando e ampliando trabalhos relacionados, restritos a pegadas digitais ativas.

1.1 Objetivos

Este trabalho explora através de algoritmos de agrupamento um conjunto de dados comportamentais e demográficos extraídos de pegadas digitais ativas e passivas de usuários de redes sociais em combinação com dados de personalidade obtidos a partir de questionários curtos de inferência de personalidade em larga escala. A exploração desses dados de alta dimensionalidade incluem a implementação de um algoritmo não supervisionado de clusterização para analisar as características presentes em cada agrupamento criado pelo algoritmo. A análise dos agrupamentos gerados é relevante para entender a viabilidade de adaptar produtos digitais, e especificamente redes sociais, para influenciar o comportamento de seus usuários, bem como suas escolhas reais.

Dessa forma, esse trabalho concentra-se em dois objetivos primários: (1) desenvolver agrupamentos criados a partir da intersecção de dados comportamentais (pegadas digitais ativas e

[1] Startup que atua especificamente no mercado de casamentos, com soluções para o planejamento e organização dos eventos

passivas) e demográficos com os traços de personalidade de usuários de redes sociais, utilizando-se o modelo dos Cinco Grandes Fatores, inferidos a partir de um questionário reduzido de autorelato, utilizando-se de algoritmos de aprendizado de máquina não supervisionados e (2) analisar qualitativamente e quantitativamente o processo de clusterização, verificando-se a criação de grupos significativos quando consideradas características socioafetivas (traços de personalidade) no agrupamento, assim como pegadas digitais passivas (navegação), a fim de entender a qualidade dos grupos formados e o quanto eles são coesos e coerentes.

Destacam-se como atividades fundamentais ao desenvolvimento desse trabalho:

- A coleta de pegadas digitais ativas e passivas de usuários em redes sociais, construindo um conjunto de dados comportamentais e demográficos de alta dimensionalidade;
- A coleta de dados de traços de personalidade no modelo dos Cinco Grande Fatores para usuários que geraram o conjunto de dados anterior, com breves questionários de inferência;
- A seleção de modelos de algoritmos não supervisionados para agrupamento de usuários de redes sociais, extraindo características da combinação dos dados coletados (comportamental, demográfico e personalidade);
- A análise e a documentação dos resultados obtidos a partir do estudo das características presentes em cada agrupamento de redes sociais;

1.2 Organização do documento

Este documento está organizado em capítulos, e cada um deles possui um assunto específico. Os capítulos 2 ao 5 correspondem ao referencial teórico, que apresentam as definições e características dos seguintes conceitos: Afeto e Computação, Pegadas Digitais e Mineração de Dados e Algoritmos de Clusterização. O capítulo 5 apresenta o atual estado da arte do estudo de traços de personalidade e redes sociais. No capítulo 6 está descrita proposta deste trabalho, com informações referentes à viabilidade da proposta, os materiais utilizados, os conjuntos de dados disponíveis e o método de pesquisa. No capítulo 7 as limitações do estudo são descritas. E no atual estágio da pesquisa é apresentado o cronograma a ser seguido para a experimentação e documentação dos resultados obtidos a partir dessa proposta.

2 AFETO E COMPUTAÇÃO

As pessoas se comportam de maneiras diferentes quando confrontadas com a mesma situação. Compreender isto é um fator-chave para entender o comportamento humano. Os ambientes computacionais que percebem essas variações no comportamento humano são qualificados para detectar e responder emoções com maior precisão (VINCIARELLI; MOHAMMADI, 2014).

Alguns fenômenos emocionais persistem por longos períodos, às vezes por toda a vida. Traços de personalidade estáveis e tendências comportamentais têm em comum um forte núcleo emocional. Isso significa que o comportamento emocional representado na ansiedade, na alegria, na hostilidade, na irritabilidade, na surpresa, no ciúme e na inveja, são exemplos de disposições emocionais. A combinação dessas disposições define a personalidade e, consequentemente, as diferenças particulares entre os seres humanos. As disposições emocionais também incluem patologias emocionais; embora estar deprimido possa ser um evento normal, a duração dele pode ser um sinal de problemas emocionais, com a intervenção médica necessária (SCHERER, 2005).

2.1 Computação Afetiva

De acordo com teorias desenvolvidas nos campos da neurociência e da psicologia, entende-se que as emoções são vistas como fundamentais para a viabilidade dos aspectos da inteligência humana no mundo real (DAMASIO, 1994). Em 1995, as pesquisas iniciadas pela pesquisadora Rosalind W. Picard introduzem o campo da Computação Afetiva (CA), englobando teorias que definem de que forma os fatores afetivos influenciam as interações entre os humanos e a tecnologia. Picard também apresenta técnicas computacionais que podem perceber e reagir às emoções humanas. Dessa forma, a Computação Afetiva destina-se ao desenvolvimento de sistemas computacionais hábeis a expressar e reconhecer estados afetivos através das interações entre humanos e Sistemas de Informação de forma natural, convincente e amigável (ZENG et al., 2007).

A Computação Afetiva é um campo de estudos multidisciplinar que engloba estudos de diversas áreas, tais como: psicologia, neurociência, ciência da computação, linguísticas e outras. Em 2010 foi lançado o periódico "*IEEE Transactions on Affective Computing (IEEE TAC)*", e desde 2005 acontece anualmente a conferência internacional da área, disseminando os resultados das pesquisas de projetos de Sistemas de Informação que podem reconhecer, interpretar e simular emoções humanas relacionadas aos fenômenos afetivos. Com objetivos claros de responder perguntas essenciais ao tempo moderno, a Computação Afetiva ajuda a entender como os estados afetivos influenciam a relação entre humanos e computadores no que diz respeito à usabilidade, percepção das emoções pelos computadores para uma melhor compreensão das pessoas, capacidade de tornar os computadores antropomórficos e questões éticas relacionadas ao empoderamento de habilidades emocionais aos computadores.

A Computação Afetiva presume que há um benefício em fornecer habilidades emocionais aos computadores, embora esse seja um debate antigo. Platão, por exemplo, argumenta que emoção e inteligência estão em lado opostos. Os Estoicos, não eram grandes fãs das emoções; Cícero descreveu que apenas é capaz de entregar-se à emoção aquele que não pode fazer nenhum uso da razão. O filósofo David Hume, por outro lado, define que a razão é, e só pode ser, escrava das emoções, e sem sentimentos faltariam motivações, impulso para agir e mesmo a razão não existiria. Dessa forma, a emoção é vista pelos filósofos como útil e que os fins são derivados de nossos desejos. Os princípios da evolução natural descritos por Darwin apontam as emoções como benéficas e úteis para a evolução da espécie humana. Por exemplo, o riso descarrega as energias acumuladas com tensões.

Todavia, a emoção não tem uma única e clara definição (ARNOLD, 1960). Na literatura filosófica e na literatura relacionada à psicologia pode-se encontrar uma diversidade de definições. O neurocientista Damasio (1994) vê a essência da emoção como uma coleção de mudanças no estado corporal que são induzidas numa miríade de órgãos por terminais nervosos e sob o controle de um sistema cerebral delicado, que está respondendo ao conteúdo de um pensamento em relação a um evento ou entidade específica. Portanto, pode-se evitar, inicialmente, discutir se os computadores podem sentir ou detectar emoções. Para isso, sugere-se um padrão de reconhecimento afetivo (PICARD, 1997).

O psicologista Ulric Neisser argumenta que os computadores não podem capturar o nível de inteligência humana porque faltam a eles corpos e emoção. O fundador-pai da área de Inteligência Artificial (IA), na mesma linha de raciocínio, responde que os sistemas inteligentes devem ter mecanismos semelhantes à emoção (SIMON, 1967). Nas últimas décadas, o interesse na relação entre emoções e personalidade em computadores foi deixado de lado, verificando-se um massivo interesse e foco da Inteligência Artificial no aspecto racional, muito mais voltado à realização eficaz de atividades do que nos aspectos da vida real, enfatizando a lógica e a racionalidade em relação às emoções.

As pesquisas publicadas por Picard, ainda em 1995, apontam uma outra direção de estudos, enfatizando as emoções como benéficas e compreendendo que os sistemas computacionais começavam a adquirir habilidades suficientes para expressar e reconhecer emoções afetivas. Assim sendo, os sistemas computacionais estavam próximos de ganhar a habilidade de “sentir emoções”.

Décadas depois da sua primeira publicação, a Computação Afetiva continua a crescer, tornando-se uma realidade e não se restringindo apenas aos esforços de pesquisas acadêmicas. Nos tempos modernos, a Computação Afetiva encontra-se em diversas aplicações. É observado casos de sucesso em novos produtos, patentes, *startups*, cursos universitários e novas iniciativas de financiamento ao redor do mundo.

Entre os seus diversos tópicos de estudos, destacam-se alguns exemplos (SCHERER; BÄNZIGER; ROESCH, 2010):

1. Reconhecimento de emoções por:

   (a) Fala: expressas nas emoções da fala natural e na detecção da depressão;

   (b) Escrita: ao analisar opiniões em mídias sociais e nos estudos da representação de expressões faciais em figuras animadas;

   (c) Faciais: a partir do estudo do impacto do envelhecimento e no reconhecimento de expressões;

   (d) Psicológicas: detecção de personalidade de usuários de redes sociais e desenvolvimento de sistemas de recomendação;

2. Síntese de emoções por:

   (a) Emoções verbais;

   (b) Emoções faciais e corporais;

3. Modelagem de emoções através de:

   (a) Influências emocionais nas tomadas de decisões;

   (b) Fatores que provocam emoções;

4. Aplicações:

   (a) Saúde: detectando doenças e modelando tratamentos;

   (b) Educação: no auxílio à aprendizagem;

   (c) Entretenimento: na detecção do estado emotivo após uma derrota ou vitória em um jogo;

   (d) Ciência comportamental.

Com o avanço de tecnologias vestíveis (*wearables*) e da facilidade na instalação dos mais variados tipos de sensores computacionais, as máquinas estão se tornando cada vez mais capazes de prever nossas reações, informando aos humanos riscos e oportunidades. Picard (1997) argumenta que, com o crescimento dos dispositivos sensoriais – capazes de monitorar, armazenar e analisar constantemente informações sobre pressão sanguínea, temperatura do corpo, nível de açúcar no sangue, tensões musculares e muitos outros estados psicológicos – aumenta-se a habilidade de mensurar e categorizar as emoções e outros estados afetivos envolvidos.

Pode-se afirmar que, sem habilidades afetivas, os computadores não são capazes de atingir um comportamento criativo e inteligente (PICARD, 1997). Percebendo-se ainda o aumento do tempo destinado à comunicação através de aparelhos computacionais, Picard (1997) enfatiza a importância da Computação Afetiva como campo de pesquisa fundamental para a evolução da Inteligência Artificial de forma crível e assertiva. Dessa maneira, pode-se concluir que a Computação Afetiva está dividida em duas ramificações de pesquisa (JAQUES, 2004):

1. Estudo da afetividade na interação humano-computador a partir da detecção e/ou expressão de emoções por computadores;

2. Investigação da simulação de componentes afetivos humanos em computadores para o desenvolvimento de sistemas mais críveis.

2.2 Personalidade

As pessoas se comportam de maneiras distintas quando confrontadas com a mesma situação. Entender as diferenças individuais são fundamentais para prever os comportamentos afetivos de cada pessoa. Computadores podem tirar vantagem dessa capacidade para melhorar a sua habilidade de perceber e responder às emoções de forma mais precisa. A personalidade é relevante para qualquer área da computação envolvendo a compreensão, previsão ou síntese do comportamento humano (VINCIARELLI; MOHAMMADI, 2014). A psicologia da personalidade é a resposta moderna para capturar características individuais estáveis e normalmente é mensurável em termos quantitativos, que explicam e preveem diferenças comportamentais observáveis entre indivíduos (VINCIARELLI; MOHAMMADI, 2014).

Um subcomponente que influencia a personalidade é a cultura. Ela consiste de regras não escritas do jogo social, sendo uma coleção de programações coletivas da mente que distingue os membros de um grupo ou a categoria de um grupo de pessoas de outros grupos de pessoas (HOFSTEDE; HOFSTEDE; MINKOV, 1991). O antropólogo francês Claude Levi-Strauss ressalta que o relativismo cultural afirma que uma cultura não tem critério absoluto para julgar as manifestações de outra cultura como “baixa” ou “nobre”. Essas manifestações da cultura ocorrem em diferentes níveis de profundidade: símbolos, heróis, rituais e valores - todos eles manifestados a partir de práticas. Hofstede, Hofstede e Minkov (1991) afirma que a cultura é oriunda do aprendizado e não é biológica: ela deriva de um ambiente social, e não dos genes herdados.

Pessoas de diferentes culturas tendem a apresentar diferentes comportamentos, motivações e emoções para a mesma situação. Entender o comportamento individual e suas motivações a cada evento experienciado também engloba o estudo da cultura. As taxas de homicídio, por exemplo, são significativamente influenciadas pela cultura; as leis e as regras não são a parte mais importante no comportamento humano (PICARD, 1997).

Na primeira metade do século 20, antropólogos sociais desenvolveram a convicção de que todas as sociedades, modernas ou tradicionais, enfrentam os mesmos problemas básicos, onde apenas as respostas entre elas diferem-se. Hofstede, Hofstede e Minkov (1991) apresentou uma análise estatística das respostas à essas questões em uma pesquisa sobre os valores culturais das pessoas em mais de 50 países ao redor do mundo. Ele definiu quatro problemas básicos que formam a Teoria das Dimensões Culturais, a qual oferece uma estrutura para examinar como os valores culturais afetam o comportamento e dá pistas sobre as formas como as pessoas de uma cultura podem agir. Por exemplo, o índice de distância do poder descreve como os membros

menos poderosos de uma sociedade aceitam e esperam certa desigualdade de poder. Dessa forma, a cultura é o ponto que existe em comum entre o homem e sua sociedade (TYLOR, 1871). Essa relação cultural e sua complexidade influenciam o homem desde o nascimento e é fator determinante na sua inclusão ao meio social e consequentemente na sua personalidade.

A personalidade é uma construção psicológica destinada a explicar a grande variedade de comportamentos humanos: únicos, estáveis e individualmente mensuráveis em termos quantitativos, para explicar e prever diferenças comportamentais (VINCIARELLI; MOHAMMADI, 2014). Testes de personalidade de autoavaliação ou classificações de observadores são sempre explorados como a verdade fundamental para testar e validar o desempenho de algoritmos de inteligência artificial para a previsão automática de tipos de personalidade. Embora exista uma grande variedade de testes de personalidade, como o Indicador do Tipo *Myers Briggs* (MBTI) (MYERS, 1962) ou o Inventário Multifásico Minnesota de Personalidade (MMPI) (GRAHAM, 1990), os instrumentos mais utilizados são os testes baseados no Modelo de Cinco Fatores. Esse modelo nasceu dos estudos sobre a teoria dos traços de personalidade, o qual descreve as dimensões humanas de forma consistente e replicável. Ele é atualmente o paradigma dominante nas pesquisas sobre personalidade e um dos modelos mais influentes de toda a psicologia (DEYOUNG; GRAY, 2009).

Considerando um amplo espectro de cenários e contextos, o estudo da conexão entre personalidade e a computação, no campo de pesquisa conhecido como Computação da Personalidade, estuda a personalidade por meio de técnicas computacionais de diferentes fontes, incluindo texto, multimídia e redes sociais. Segundo os autores Vinciarelli e Mohammadi (2014), a Computação da Personalidade aborda três problemas fundamentais:

1. Reconhecimento Automático de Personalidade (APR);
2. Percepção Automática de Personalidade (APP);
3. Síntese Automática de Personalidade (APS).

Especificamente sobre o Reconhecimento Automático de Personalidade, trabalhos científicos demonstraram a validade de soluções a partir da coleta de diferentes pegadas digitais, em particular as preferências do usuário, como imagens de perfil do Facebook (SEGALIN et al., 2017), e mostraram que as máquinas podem reconhecer a personalidade melhor do que os seres humanos (YOUYOU; KOSINSKI; STILLWELL, 2015). Pode-se caracterizar esta pesquisa, no escopo da Computação da Personalidade, dentro do estudo da aplicabilidade e do potencial do uso da detecção da personalidade, a partir do Reconhecimento Automático de Personalidade, para o desenvolvimento de aplicações em diversas áreas que serão estudadas no capítulo 5.

2.3 Modelo dos Cinco Grande Fatores

O Modelo dos Cinco Grandes Fatores, embora tenha sofrido tentativas de ser enriquecido com mais dimensões, se manteve estável ao longo dos estudos. Tornou-se um modelo ampla-

mente utilizado por cientistas, definindo as características de cinco traços de personalidade:

1. **Abertura à Experiência**: A abertura está relacionada à imaginação, criatividade, curiosidade, tolerância, liberalismo político e apreciação pela cultura. Pessoas que pontuam alto neste traço apreciam ideias novas e incomuns e têm um bom senso estético;

2. **Conscienciosidade**: A conscienciosidade mede a preferência por uma abordagem organizada da vida, em contraste com uma abordagem espontânea. As pessoas com alta pontuação em conscienciosidade têm maior probabilidade de serem bem organizadas, confiáveis e consistentes. Elas gostam de planejar, buscam conquistas e perseguem metas de longo prazo. Indivíduos não conscientes são geralmente mais leves, espontâneos e criativos. Eles tendem a ser mais tolerantes e menos sujeitos a regras e planos;

3. **Extroversão** A extroversão mede uma tendência a buscar estímulo no mundo externo, a companhia de outros e a expressar emoções positivas. Pessoas com alta pontuação neste traço tendem a ser mais extrovertidas, amigáveis e socialmente ativas. Elas são geralmente enérgicas e falantes; não se importam de estar no centro das atenções e fazem novos amigos mais facilmente. É mais provável que os introvertidos sejam solitários ou reservados e busquem ambientes caracterizados por níveis mais baixos de estímulo externo;

4. **Agradabilidade**: A agradabilidade se refere a um foco em manter relações sociais positivas, ser amigável, compassivo e cooperativo. As pessoas com uma pontuação alta tendem a confiar nos outros e adaptar-se às suas necessidades. Por outro lado, pessoas com níveis mais baixos são mais focadas em si mesmas, menos propensas a se comprometer e podem ser menos ingênuas. Elas também tendem a ser menos vinculadas às expectativas e convenções sociais;

5. **Estabilidade Emocional** A estabilidade emocional, inversamente chamada de **Neuroticismo**, mede a tendência de experimentar mudanças de humor e emoções como culpa, raiva, ansiedade e depressão. Pessoas com baixa pontuação de estabilidade emocional (alto neuroticismo) têm maior probabilidade de experimentar estresse e nervosismo, enquanto pessoas com alta pontuação (baixo neuroticismo) tendem a ser mais calmas e autoconfiantes.

A Computação Afetiva é fortemente impactada pelo estudo da personalidade. A fim de tornar computadores capazes de responder às diferenças humanas, originadas da cultura e da personalidade, desenvolveu o campo da Computação da Personalidade. Com um crescimento de interesse constante a partir do ano de 2006, ela objetiva a resolução de três problemas computacionais: o reconhecimento verdadeiro da personalidade de um indivíduo, a predição da personalidade de um indivíduo e a geração de personalidades artificiais através de agentes conversacionais (VINCIARELLI; MOHAMMADI, 2014).

Um dos mais maiores desafios para a computação é lidar com os humanos. Não apenas no aspecto restrito da interação direta entre humanos e computadores, mas nas suas diversas formas, como usuários, como dados a serem analisados ou mesmo como consumidores de materiais criados artificialmente. A personalidade, como um construto capaz de capturar os aspectos únicos de um indivíduo, deve se tornar a chave para melhor preencher a lacuna entre pessoas e computadores. No entanto, o tamanho das amostras de estudos de personalidade é geralmente pequena para validação estatística e com muitos preconceitos. Por exemplo: pessoas brancas, educadas, industrializadas, ricas e democráticas (LAMBIOTTE; KOSINSKI, 2014).

3 PEGADAS DIGITAIS

Devido ao crescimento das redes sociais online, a quantidade de pegadas digitais dos usuários está aumentando exponencialmente. Diariamente as pessoas estão experimentando interações sociais, entretenimento e atividades da vida em geral em serviços online mediados por dispositivos digitais. Todos esses registros são conhecidos como *Big Social Data* (LAMBIOTTE; KOSINSKI, 2014) – referindo-se a grandes volumes de dados que se relacionam às pessoas, descrevendo seu comportamento e suas interações sociais mediadas pela tecnologia no mundo digital (OLSHANNIKOVA et al., 2017). Ao contrário de pegadas físicas, os rastros digitais são normalmente permanentes e indeléveis, com uma enorme quantidade de dados pessoais valiosos. O valor por trás das pegadas digitais está nos registros minuciosos e detalhados do comportamento humano e em suas interações sociais muito específicas (GOLDER; MACY, 2014). Por exemplo, todos os dias milhões de pessoas em todo o mundo expressam seus pensamentos, emoções e crenças imediatas escrevendo, publicando e compartilhando conteúdo nas mídias sociais ou mesmo navegando em conteúdos públicos.

Antes da revolução das mídias sociais, embora já houvesse abundância em teorias sociais, a coleta de dados sociais era uma tarefa complexa e de alto custo, pois observar a vida social é difícil e desgastante (GOLDER; MACY, 2014). As mídias sociais revolucionaram a maneira como as pessoas interagem umas com as outras, ao mesmo tempo em que propicia as oportunidades de pesquisa para abordar questões sobre comportamento social a partir da mineração de dados de pegadas digitais deixadas pelos usuários, criando assim novas metodologias de pesquisa que não dependem exclusivamente de dados observacionais coletados a partir de experimentos de campo. Os registros dessas atividades estão mudando o paradigma na pesquisa em ciências sociais, passando de estudos de pequena escala e longos para estudos de larga escala e rápidos.

Existem evidências de pesquisa sobre a eficácia de prever traços psicológicos de suas pegadas digitais (LAMBIOTTE; KOSINSKI, 2014). Dessa forma, a análise das pegadas digitais está mostrando não apenas a identidade online dos usuários, mas como registros generalizados das pegadas digitais podem ser usados para inferir a personalidade. Mesmo que os usuários não estejam produzindo pegadas digitais ativamente, suas conexões virtuais, como amigos e familiares, estão produzindo passivamente pegadas para os mesmos usuários.

Pode-se concluir que pegadas digitais são dados sociais criados por usuários quando eles interagem com os canais de mídia. Essas pegadas digitais não são apenas identidades, mas também memórias, momentos e comportamentos. Provedores de mídia social que coletam essas enormes crônicas digitais podem determinar como e por que os usuários se comportam e compram em plataformas digitais (FISH, 2009). Além disso, esses rastros digitais ajudam as empresas a analisar os sentimentos e o perfil psicológico dos clientes usando análises avançadas para obter uma percepção mais profunda de seu comportamento e perfil (CHARLESWORTH, 2014).

O futuro da aplicabilidade e das questões éticas do rastreamento de pegadas digitais pode ser muito complexo e amplo. Os autores Arakerimath e Gupta (2015) analisaram os pontos positivos e negativos dessa técnica, argumentando que as pegadas digitais trazem benefícios aos usuários de produtos digitais, tais como a recomendação de conteúdo e produtos. E, também, às empresas que podem a partir da mineração de dados extrair atributos importantes das pegadas digitais para diversos fins. Uma delas pode ser a prévia avaliação de um candidato à uma nova oportunidade – antes mesmo de uma entrevista. Além disso, sites de relacionamento podem utilizar apenas as pegadas digitais passivas (mais informações na Seção 3.2) para criar algoritmos que potencializam combinações entre os gostos e desprazeres das pessoas presentes em redes complexas. Os autores também afirmaram que, com o potencial de informação valiosa presentes nas pegadas digitais, novas oportunidades de negócios poderão ser desenvolvidas, como a criação de serviços específicos de gestão de pegada digitais, desde o nascimento de uma pessoa até a morte (embora, ainda seja discutível a temporariedade desses dados, pois as pegadas digitais podem estar armazenadas para sempre). Todo esse potencial, segundo os autores, evidenciam pontos negativos nessa técnica. Por exemplo, as pegadas digitais podem ajudar empresas ou mesmo outras pessoas a prever nossos traços psicológicos íntimos – mesmo que as pessoas não desejem compartilha-los – utilizando-os para diversos fins, como na política em geral.

3.1 Pegadas Digitais Ativas

É difícil, senão impossível, existir na sociedade contemporânea sem deixar vestígios digitais. Publicações em redes sociais, e-mails enviados, transações digitais, conversas públicas e privadas, compartilhamento de localização em tempo real, álbuns de fotos digitais, comentários em publicações de mídia e muitas outras informações são coletadas e mantidas em uma ampla variedade de locais, sob o controle de um grande variedade de entidades e armazenadas por períodos indefinidos de tempo. Todas essas atividades digitais são exemplos de pegadas digitais ativas.

As pegadas digitais ativas são criadas quando um usuário compartilha seus dados com consentimento, ou seja, o usuário têm o conhecimento prévio de que os seus dados poderão ser utilizados por empresas e terceiros (ARAKERIMATH; GUPTA, 2015). Por exemplo, em um ambiente online, quando um usuário cria um perfil de rede social ou comenta alguma postagem ou artigo, ele está criando uma pegada digital ativa de si mesmo. Com isso, grandes plataformas comerciais de redes sociais que armazenam pegadas digitais usufruem e criam uma oportunidade sem precedentes de observar pessoas em ambientes realistas. A pesquisa de Kosinski et al. (2015a), que estuda a rede social Facebook como ferramenta para o estudo de Ciências Sociais, enumera as principais pegadas digitais ativas presentes na rede social, tais como:

- **Perfil demográfico:** Composto por um ID de usuário exclusivo, nome completo, foto do perfil, idade, sexo, status de relacionamento, interesses românticos, localização geográ-

fica, local de origem, histórico de trabalho e educação, biografia, link para site pessoal, fuso horário, política e religião pontos de vista, interesses gerai; e listas de músicas, filmes, programas de TV, livros, citações e esportes favoritos;

- **Conteúdo gerado pelo usuário:** Consiste em atualizações de status, fotos, vídeos, comentários no conteúdo ou páginas de outras pessoas, links e notas publicados por usuários ou amigos. Cada parte do conteúdo também contém metadados, como as posições das pessoas presentes na imagem, data de publicação, lista de pessoas que curtiram, suas configurações de privacidade, etc;

- **Estrutura da rede social:** Contém a lista de amigos e o tipo de conexões de usuários. Os tipos de conexão incluem amizades, vínculos familiares e seguidores.

- **Preferências e atividades do usuário:** Incluindo curtidas, participações em grupos e eventos, aplicativos instalados e *tags* em fotos ou postagens;

- **Informações sobre os amigos dos usuários:** Detalhes demográficos e atividades de amigos visíveis para um determinado usuário;

- **Mensagens privadas entre usuários:** Geralmente escritas e enviadas através do recurso de mensagens instantâneas.

Todo esse conjunto de pegadas digitais ativas também podem ser seletivamente modificadas pelos usuários, da mesma forma que ocorre com questionários de auto-avaliação. Ou seja, os dados podem ser afetados por vieses induzidos pelo usuário, típicos dos autorrelatos, como a conveniência social e deturpação intencional. Entretanto, de acordo com a pesquisa de Back et al. (2010) os perfis do refletem o perfil de personalidade real e não auto-idealizado de seus proprietários. Kosinski, Stillwell e Graepel (2013) mostraram que outros elementos do Facebook (como sexo, idade, pontos de vista políticos, religião e curtidas do Facebook) inferem relacionamentos consistentes e significativos nos perfis. Quando analisa-se a extração de pegadas digitais para outros fins, também encontra-se resultados relevantes. O estudo de Önder, Koerbitz e Hubmann-Haidvogel (2016), a partir de pegadas digitais ativas de usuários que compartilharam fotos com identificação geográfica em um aplicativo de compartilhamento de imagens como fotografias chamado de *Flickr*, discute indicadores da demanda turística a partir desse rastro digital. Os resultados do estudo confirmaram que esses dados podem ser usados para elaboração de ferramentas de estimativa de turistas nas cidades austríacas.

Embora o conjunto de pegadas digitais ativas seja grande, apenas uma dimensão desses dados aplicados a algoritmos de aprendizado de máquina pode oferecer resultados de inferência de perfil de um usuário de rede social mais assertivos que o julgamento humano ao analisar traços íntimos. Estudos em tópicos avançados de *deep learning*, como a pesquisa de Wang e Kosinski (2018), mostra que os rostos contêm muito mais informações sobre orientação sexual do que podem ser percebidas e interpretadas pelo cérebro humano. Ao extrair características

de um conjunto de dados de 35.326 imagens faciais e com apenas uma foto por perfil, o algoritmo distinguiu corretamente entre homens gays e heterossexuais em 81% dos casos e em 71% dos casos para mulheres, enquanto os juízes humanos alcançaram uma precisão muito menor: 61% para homens e 54% para mulheres. A precisão do algoritmo aumentou para 91% e 83%, respectivamente, quando analisado perfis com cinco imagens faciais por pessoa.

3.2 Pegadas Digitais Passivas

Os usuários de Internet e, especificamente, os usuários excessivos de mídias sociais, podem não estar cientes de todas as pegadas digitais deixadas por eles nesses ambientes. Os serviços das principais plataformas de mídia social redefiniram as maneiras pelas quais seus negócios geram valor. Com o rastreamento massivo e, potencialmente, invasivo e onipresente, eles usam algoritmos para gerar informações poderosas por meio de conexões, inferências e interpretações de dados (DWORK; MULLIGAN, 2013). Essa questão é ainda mais relevante quando trata-se de pegadas digitais passivas. Isso se deve ao fato de que as pegadas digitais não são apenas o produto da participação ativa por meio da produção e compartilhamento de conteúdo, mas também podem ser geradas por algoritmos, por outros usuários da Internet ou de forma inconsciente por uma pessoa. Portanto, as pegadas digitais são a soma dos dados produzidos, tanto por formas ativas quanto por formas passivas de participação (LUTZ; HOFFMANN, 2017).

As pegadas digitais passivas são o conjunto de rastros digitais que as pessoas deixam online de forma não intencional (FISH, 2009). Quando você visita um site, ele pode registrar seu endereço IP, que identifica seu provedor de serviços de Internet e sua localização aproximada. Por exemplo, sites que coletam informações sobre a frequência de uso e o conteúdo consumido por um usuário de rede social estão adicionando ao seu banco de dados pegadas digitais deixadas de forma passiva. Além disso, a coleta de pegadas digitais passivas não está exclusivamente vinculada à navegação na Internet. Ou seja, usuários podem não estar cientes de que suas informações digitais estão sendo coletadas em grande escala a partir de dispositivos como televisores e carros inteligentes, câmeras e demais sensores sensores inteligentes (WILLIAMS; PENNINGTON, 2018). Outro exemplo de pegadas digitais passivas são aquelas criadas de forma ativa por outras pessoas a respeito de outras. Os autores Ophir, Asterhan e Schwarz (2019) que estudaram a relação entre as pegadas digitais e a depressão adolescente, a rejeição social e a vitimização do bullying no Facebook, descrevem como exemplo desse tipo de dado postagens no Facebook que mencionam outras pessoas, podendo esse conteúdo ser textual ou de mídia como as fotos compartilhadas que podem incluir marcações de terceiros.

4 MINERAÇÃO DE DADOS E ALGORITMOS DE CLUSTERIZAÇÃO

Como foi observado nas seções anteriores, as redes sociais são exemplos de fontes de grandes conjuntos heterogêneos de dados. Esses dados criam a oportunidade de análise de redes sociais que visa facilitar computacionalmente os estudos sociais e as relações humano-sociais (JIANG; LEUNG; PAZDOR, 2016). Dessa forma, a mineração de dados e, especificamente, os algoritmos de clusterização ou agrupamento de dados, são cada vez mais aplicados nessas áreas de estudo.

A mineração de dados é um termo amplo, usado para descrever diferentes aspectos do estudo da coleta, limpeza, processamento, análise e obtenção de percepções de dados (AGGARWAL, 2015). A partir da mineração de dados, a descoberta de estruturas interessantes, inesperadas ou valiosas em grandes conjuntos de dados é uma atividade viável. Como tal, tem dois aspectos distintos. O primeiro diz respeito a estruturas "globais"em larga escala com o objetivo de modelar as suas formas, ou as características das suas formas e suas distribuições. O segundo diz respeito à pequena escala "local"de estruturas, onde o objetivo é detectar anomalias e decidir se elas são reais ou ocorrências aleatórias. Para englobar esses dois aspectos, a mineração de dados moderna combina estatística com ideias, ferramentas e métodos da ciência da computação, aprendizado de máquina, tecnologias de banco de dados e outras tecnologias clássicas de análise de dados (HAND, 2007).

Hand (2007) ainda define que a mineração de dados pode ser divida em dois grupos de ferramentas. O primeiro é sobre o desenvolvimento de modelos, o qual é responsável pelo desenvolvimento de resumos descritivos em alto nível de conjuntos de dados, utilizando-se de estatística moderna que incluí modelos de regressão, decomposição de *cluster* e Redes Bayesianas. São esses modelos que descrevem de forma geral cada conjunto de dados analisados. Por outro lado a descoberta de padrões está diretamente relacionada a entender estruturas de dados locais, em um vasto espaço de dados, descrevendo dados com alta densidade de anomalia quando comparada com a *baseline* original. Embora ambos trabalhem diretamente com dados, o segundo grupo de ferramentas de mineração de dados (descoberta de padrões) necessita fundamentalmente de qualidade nos dados presentes em cada conjunto analisado.

À medida que o uso massivo de mídias sociais cresce, também aumenta a mineração de dados de mídias sociais para obter conhecimento nas opiniões, humor, redes e relacionamentos dos usuários comuns de mídia social e principais influenciadores. Segundo os autores Kennedy, Elgesem e Miguel (2017), a mineração de dados de mídia social inclui uma ampla gama de atividades realizadas para analisar, organizar, classificar e entender esses dados, desde contar gostos e compartilhamentos de conteúdo até medir alcance, sentimentos e principais influenciadores. Porém, os dados de mídia social têm três características que colocam desafios para os pesquisadores: os dados são grandes, ruidosos e dinâmicos (BARBIER; LIU, 2011). Para superar isso, utilizam-se diversas técnicas como análise de redes sociais, processamento de linguagem natural e demais algoritmos de aprendizado de máquina (KENNEDY; ELGESEM; MIGUEL,

2017).

Os algoritmos supervisionados e não supervisionados são usados para identificar os padrões ocultos nos dados. Os autores Barbier e Liu (2011), em seus estudos de mineração de dados aplicado às redes sociais, definem que as abordagens supervisionadas dependem de algum conhecimento a priori dos dados (por exemplo, rótulos de classe). Em contrapartida, algoritmos não supervisionados são usados para caracterizar dados sem nenhuma instrução prévia sobre quais tipos de padrões serão descobertos pelo algoritmo. Embora exista uma variedade de trabalhos realizados até o momento referentes à mineração de dados de dados em redes sociais, utilizando ambos os tipos de algoritmos de aprendizado (supervisionado ou não supervisionado), a abordagem escolhida depende do conjunto de dados (dados com rótulos, dados sem rótulos e dados com apenas uma pequena porção de rótulos) e da pergunta específica que está sendo investigada (BARBIER; LIU, 2011).

Como essa pesquisa estuda como agrupar dados de personalidade, comportamento e demografia sem conhecer os rótulos de cada potencial agrupamento, o estudo de algoritmos não supervisionados são fundamentais para o estudo. O *clustering* é uma técnica comum de mineração de dados não supervisionada que é útil ao confrontar conjuntos de dados sem rótulos (BARBIER; LIU, 2011). Dutt, Ismail e Herawan (2017) realizaram a revisão sistemática sobre mineração de dados no âmbito educacional, definindo a clusterização como uma técnica para coletar e apresentar itens de dados similares e questionando o que de fato define similaridade, para em seguida afirmar que essa é a questão chave para o entendimento de "agrupamento". Ainda segundo os autores, na notação estatística, o agrupamento é o algoritmo não supervisionado de aprendizagem mais importante. Para Berkhin (2006), o objetivo principal de um algoritmo de clusterização é criar uma divisão de dados em grupos de objetos semelhantes e generalizar padrões ocultos para a representação de conceitos distintos de dados. Isso significa, de uma perspectiva prática, que o *clustering* desempenha um papel destacado em aplicativos de mineração de dados, como exploração científica de dados, recuperação de informações e mineração de texto, aplicativos de banco de dados espaciais, análise da Web, CRM, marketing, diagnóstico médico, biologia computacional e muitos outros (BERKHIN, 2006). Dessa forma, os algoritmos de clusterização ou agrupamento de dados determinam quais elementos no conjunto de dados são semelhantes entre si com base na similaridade dos elementos de dados, onde essa semelhança pode ser definida como distância euclidiana para alguns conjuntos de dados numéricos. Entretanto, nos dados associados às mídias sociais, as técnicas de cluster geralmente devem ser capazes de lidar com o texto, onde as técnicas de agrupamento usam palavras-chave representadas como um vetor e a medida de similaridade de cosseno é usada para distinguir quão semelhante um vetor é em relação a outro (BARBIER; LIU, 2011).

Os autores Berkhin (2006) categorizam os algoritmos de clusterização em diversos tipos como hierárquicos, particionáveis, baseados em *grid*, baseados em coocorrência de dados categóricos, baseados em restrições e muitos outros. Porém, tradicionalmente, as técnicas de agrupamento são normalmente divididas em hierárquicas e particionáveis. Enquanto os algo-

ritmos hierárquicos criam agrupamentos gradualmente, os algoritmos de particionamento (ou *clustering*) criam os agrupamentos diretamente. Ao fazer isso, eles tentam definir agrupamentos realocando iterativamente pontos entre subconjuntos ou tentam identificar agrupamentos como áreas altamente preenchidas com dados (BERKHIN, 2006). Em uma perspectiva prática, a clusterização hierárquica se concentra em quão bem os pontos se encaixam em seus agrupamentos e tendem a criar agrupamentos com formas convexas adequadas. No caso dos algoritmos de particionamento, eles se concentram em descobrir conjuntos densos de dados conectados, mas flexíveis em termos de formato.

Dessa maneira, para o objetivo dessa pesquisa, os algoritmos de particionamento que dividem os dados em vários subconjuntos e os melhoram gradualmente para obtenção de grupos de alta qualidade são os indicados para esse trabalho. Essa indicação ocorre pois essa técnica possibilita obter informações valiosas do conjunto de dados explorado, analisando quais padrões comportamentais e demográficos correlacionam-se aos traços de personalidade, e consequentemente, permitindo quais características são mais relevantes a cada traço de personalidade de forma não excludente. Ou seja, não restringindo a análise comportamental e demográfico apenas ao traço de personalidade mais proeminente de cada usuário, e sim ao conjunto de traços que formalizam a personalidade deles no modelo dos Cinco Grandes Fatores. Dessa forma, entre os muitos algoritmos de particionamento de dados, os autores Khawaja et al. (2016) definem que o algoritmo de *K-means* é um dos mais aplicados no campo de mineração de dados e aprendizado de máquina para esse fim. De acordo com Hartigan e Wong (1979), ele foi descrito com detalhes em 1975 por Hartigan e tem como objetivo dividir os pontos M nas dimensões N em *clusters* K, para que a soma dos quadrados dentro do *cluster* seja minimizada. Em uma perspectiva prática, isso significa encontrar uma alocação dos dados em grupos distintos (*clusters*) de maneira que, dentro de cada *cluster*, os dados estejam o mais próximos possível. Isso assume que, quanto mais próximos eles estiverem, mais parecidos eles serão. A aplicabilidade dessa técnica é discutida em diversos exemplos citados como respostas às questões principais de pesquisa no capítulo 5.

4.1 Algoritmos de Clusterização Selecionados

Com dados brutos acessíveis com alta dimensionalidade, e, consequentemente, prováveis altos índices de ruído, como aqueles que compõem o conjunto de dados proposto na pesquisa, e revisadas na Seção 6.4, sugere-se, potencialmente, alguns algoritmos específicos de aprendizado de máquina não supervisionado, para potencializar a análise exploratória dos dados coletados, após etapas prévias de pré-processamento. Dessa forma, com o objetivo de encontrar grupos semelhantes a partir das amostras coletadas e revelar padrões no conjunto de dados, essa seção introduz alguns algoritmos indicados para a pesquisa.

Entre os algoritmos sugeridos a seguir, todos eles requerem uma prévia configuração sobre o número de clusters (denominados de k), embora nem sempre seja claro qual o valor ideal a

ser definido (HAMERLY; ELKAN, 2004). Embora esse seja um fator que facilite a comparação entre distintos algoritmos escolhidos para uma tarefa, como a do trabalho aqui proposto, de acordo com Hamerly e Elkan (2004), a correta escolha de k fica mais difícil quando os dados têm muitas dimensões, mesmo quando os clusters estão bem separados. Embora algumas técnicas avançadas sejam sugeridas para o trabalho com matrizes esparsas, pesquisa mostram que pode ser preferível usar técnicas de redução de dimensionalidade antes do agrupamento e, em seguida, verificar as melhoras formações por algoritmos que atuem a partir de dados com dimensionalidade reduzidas (DASGUPTA, 2013; NIU; DY; JORDAN, 2011; COHEN et al., 2015; BOUTSIDIS et al., 2014). Abaixo são citados os algoritmos de clusterização testados nesse trabalho.

- **K-means**: Relativamente simples de implementar, algumas das vantagens desse algoritmo são que ele é escalável para grandes conjuntos de dados, possui convergência garantida e facilidade de adaptação a novas amostras, segundo os autores Santini (2016); Li e Wu (2012); Celebi, Kingravi e Vela (2013). Eles também citam como desvantagens do algoritmo a dependência explícita na escolha de valores iniciais para definição dos melhores valores dos centroides[1], dificuldades para agrupar dados de tamanhos e densidades variados, e *outliers* podem obter seu próprio cluster ao invés de serem ignorados pelo algoritmo. Previamente introduzido na seção anterior, o algoritmo de K-means busca encontrar *k* divisões para satisfazer um critério objetivo, onde primeiramente é escolhido alguns pontos para representar os pontos focais do cluster inicial e posteriormente, reúne-se pontos de amostra restantes em seus pontos focais de acordo a distância euclidiana (LI; WU, 2012). Ou seja, esse algoritmo agrupa os dados tentando separar amostras em grupos de variância similar, minimizando um critério conhecido como inércia ou soma dos quadrados dentro do cluster;

- **K-medoids**: O algoritmo K-medoids é usado para encontrar medoids[2] em um cluster que é um ponto localizado no centro de um cluster (ARORA; VARSHNEY et al., 2016). Como uma alternativa ao K-means, que é sensível a amostras com valores muito distantes da distribuição normal do conjunto, ocasionando distorção na criação de agrupamentos, o K-medoids busca minimizar a soma de dissimilaridades entre objetos rotulados para estar em um mesmo cluster e um dos objetos designados como representante desse cluster. Com isso, quando há a necessidade do desenvolvimento de clusters resilientes a *outliers*, o K-medoids pode ser uma alternativa relevante a ser validada;

- **Spectral Clustering**: Assim como os demais algoritmos selecionados, o Spectral Clustering é simples de implementar, e muitas vezes supera em desempenho os algoritmos de

[1] Centroide é o ponto associado a uma forma geométrica, também conhecida como centro geométrico.

[2] Medoids são objetos representativos de um cluster dentro de um conjunto de dados cuja dissimilaridade média com todos os objetos no cluster é mínima.

clusterização tradicionais. O algoritmo baseia-se nos princípios de vetores de Eigen [3] de alguma matriz com base na distância entre os pontos (ou outras propriedades) e, em seguida, usá-os para agrupar os vários pontos (VERMA; MEILA, 2003). Ou seja, enquanto o algoritmo de K-means preocupa-se com a distância euclidiana,o Spectral Clustering, concentra-se na conectividade por ser semi-convexo, reduzindo conjuntos de dados multidimensionais complexos em clusters de dados semelhantes em dimensões mais raras.

- **Agglomerative Clustering**: O Agglomerative Clustering é um agrupamento hierárquico importante e bem estabelecido em aprendizado de máquina não supervisionado. Ele baseia-se em um processo de cluster de baixo para cima, ou seja, inicialmente cada exemplo de entrada forma seu próprio cluster e em passos subsequentes, os dois clusters mais próximos são mesclados até que apenas um cluster permaneça (ACKERMANN et al., 2014). O resultado é uma representação baseada em árvore dos objetos, chamada de dendrograma [4]. Por sua representação visual, em comparação com o K-means, ser mais informativa do que um conjunto não estruturado de clusters plano, torna mais fácil decidir sobre o número de clusters ideal para a análise desejada.

[3] Um vetor Eigen é um vetor diferente de zero que é mapeado por uma dada transformação linear de um espaço vetorial.

[4] Um dendrograma é um tipo específico de diagrama ou representação icônica que organiza determinados fatores e variáveis em uma árvore.

5 TRABALHOS RELACIONADOS

Esse capítulo tem como objetivo descrever o estado da arte nas principais disciplinas que realizam o estudo combinatório do comportamento humano, da personalidade e das redes sociais. Para formular as perguntas centrais do estudo realizado nessa pesquisa, bem como seus objetivos, foi realizada uma busca baseada em combinações de palavras-chaves nos principais canais de divulgação de estudos na área.

A definição das (1) perguntas centrais e a (2) escolha das palavras-chave para estudos primários relevante, são passos iniciais em um processo de revisão sistemática. Petersen et al. (2008), define que, além dessas duas etapas iniciais, outras três etapas devem ser seguidas posteriormente: (3) triagem dos documentos, (4) *keywording* dos resumos e (5) extração de dados e mapeamento. Dessa forma, dado que as questões de pesquisa devem exemplificar os objetivos do estudo de mapeamento, as principais questões foram elaboradas:

QP$_1$: Qual o impacto de traços de personalidade no comportamento de um usuário de redes sociais?

QP$_2$: Qual a relação entre pegadas digitais em redes sociais e a personalidade de seus usuários?

QP$_3$: Quais métodos/técnicas de mineração e aprendizado de máquina são empregadas para detectar automaticamente a personalidade de usuários de redes sociais?

QP$_4$: Quais métodos/técnicas de mineração e aprendizado de máquina são empregadas para agrupar usuários de redes sociais de acordo com seu comportamento e personalidade?

QP$_5$: Qual a dimensionalidade[1] de dados utilizados nas pesquisas de agrupamento e detecção de personalidade de usuários em redes sociais?

QP$_6$: A detecção da personalidade pode ser utilizada para influenciar o comportamento de usuários nas redes sociais? Como?

Para o entendimento do potencial de pesquisa na área de personalidade e redes sociais, uma questão secundária também foi definida:

QS$_1$: Quais são as questões modernas relacionadas à privacidade no uso de pegadas digitais ativas e passivas para objetivos comerciais e de pesquisa?

O seguinte conjunto de palavras-chave foi utilizado para a busca de trabalhos relacionados[2]: *personality AND social networks*, *clustering personality*, *digital footprints AND social networks*, *digital, footprints AND personality*, *social network behaviour*, *social network and clustering* e *personality computing*. Essa busca foi realizada nas bases da ACM, IEEE e Springer, com o Google Scholar sendo utilizado de forma complementar.

Com buscas sequenciais realizadas nas bases anteriormente citadas, entre todos os artigos encontrados, um total de 111 artigos foram escolhidos a partir de seus títulos e por sua relação

[1] Dimensionalidade se refere a quantos atributos um conjunto de dados possui. Um conjunto de dados com alta dimensionalidade é um conjunto com muitos atributos. Dessa forma, assim como na estatística, a dimensionalidade se refere a quantos atributos um conjunto de dados possui.

[2] Os termos utilizados para a busca de trabalhos relacionados também foram traduzidos para o português e submetidos ao agregador Google Scholar, embora um volume baixo de resultados foi encontrado

com as questões centrais da pesquisa. Após a leitura dos resumos desses artigos, apenas os trabalhos que visavam a detecção e o agrupamento de traços de personalidade, bem como o estudo de padrões comportamentais de acordo com traços de personalidade em redes sociais, foram selecionados. Essa seleção inicial resultou em 58 artigos (52% do total de artigos relacionados), como pode ser verificado na Figura 1.

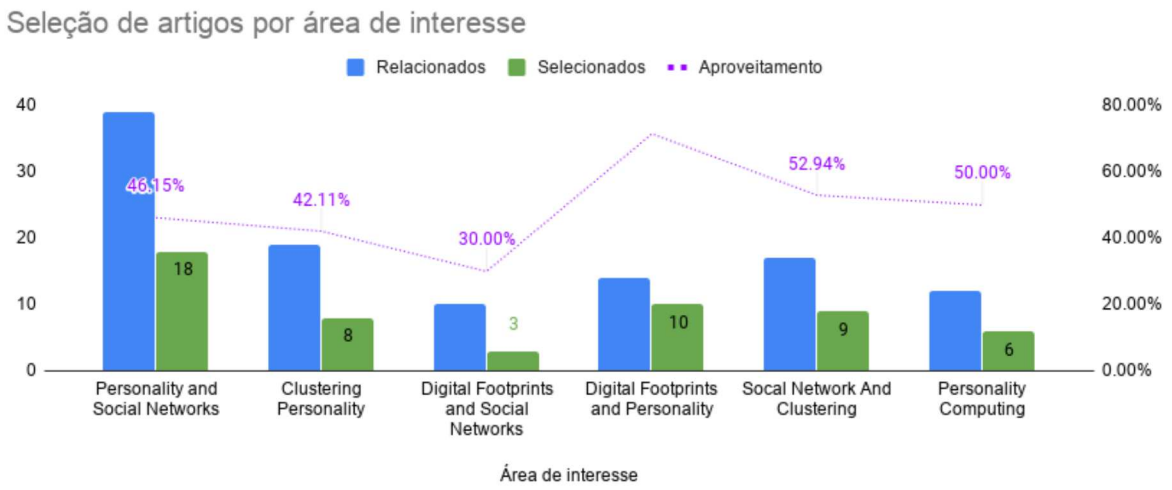

Figura 1: Atualização de trabalhos pré-selecionados por área de interesse

Na Figura 2, que ilustra a distribuição de artigos selecionados por título e por resumo nas bases previamente citadas, observa-se uma amostra baixa na área de interesse "Agrupamento de Personalidade". Artigos dessa área de interesse são relevantes para o correto entendimento e aplicabilidade das melhores práticas, métodos e processos utilizados para agrupamento de características relacionadas à personalidade em diferentes segmentos de estudo. A premissa de concentrar a pesquisa no estado da arte mais recente foi relevante para priorizar trabalhos contemporâneos para essa análise. Aproximadamente 70% dos trabalhos selecionados foram publicados nos últimos três anos, como pode ser visto na Figura 3.

A terceira etapa de seleção de trabalhos relacionados contou com a leitura completa dos artigos pré-selecionados na etapa anterior, buscando-se priorizar artigos recentes e correlacionados às questões principais da pesquisa proposta. Nessa análise foram desconsiderados os artigos *short-paper* ou que não fizeram uma avaliação devido à falta de informações mais completas e resultados sobre a abordagem proposta. Dessa forma, 22 artigos foram selecionados para a etapa final (20% dos artigos selecionados por título e 38% dos selecionados pelo resumo).

Na quarta e última etapa os artigos selecionados foram revisitados, a fim de extrair as informações necessárias para responder às questões principais e secundárias de pesquisa definidas. A lista completa dos trabalhos relacionados selecionados está disponível na Tabela 1.

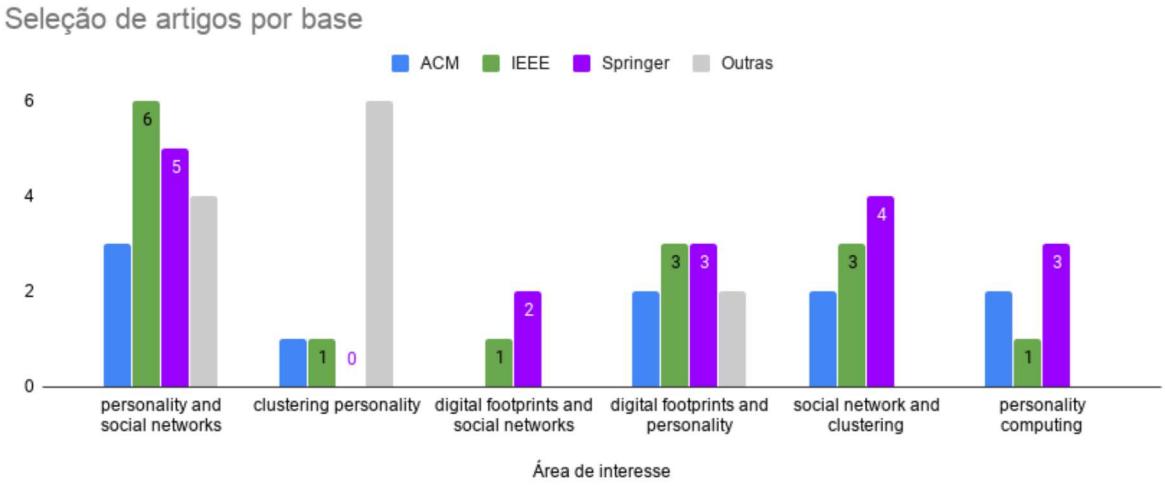

Figura 2: Distribuição dos artigos selecionados por base

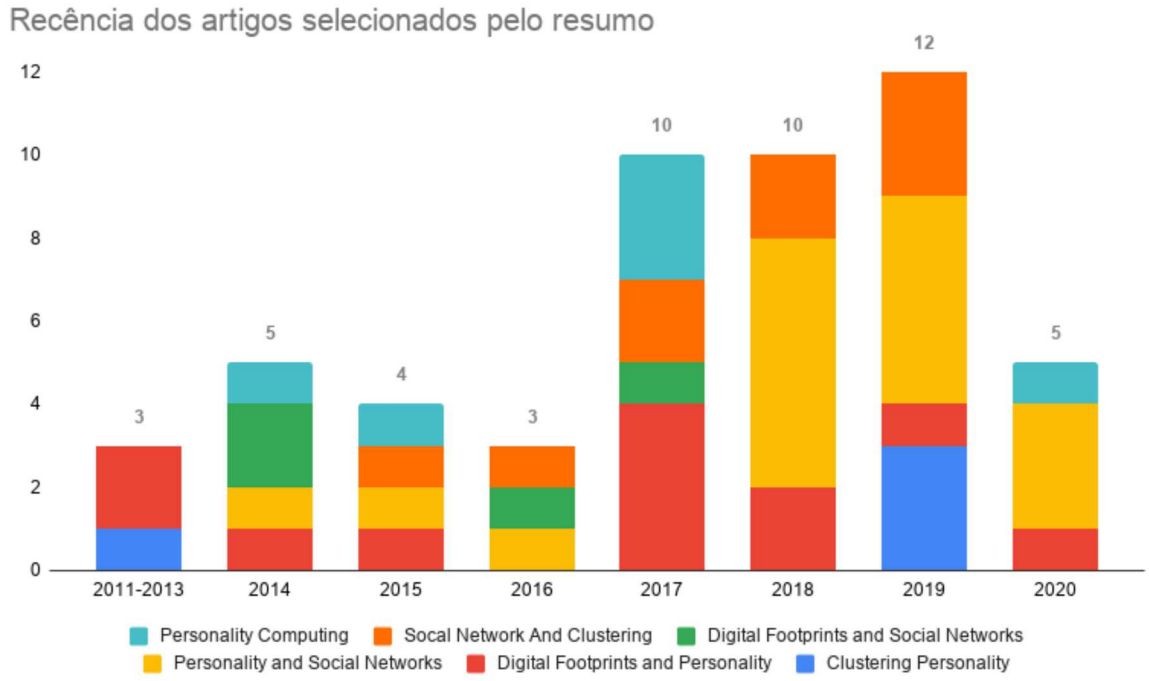

Figura 3: Recência dos trabalhos selecionados para análise

Tabela 1: Trabalhos relacionados selecionados

Artigo	**Autores**	**Ano**	**Base**
The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services	Obar, Jonathan A e Oeldorf-Hirsch, Anne	2020	Information, Communication & Society
Clustering based personality prediction on turkish tweets	Tutaysalgir, Esen e Karagoz, Pinar e Toroslu, Ismail H	2019	Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
Using Consumers' Digital Footprints for More Persuasive Mass Communication	Matz, Sandra e Kosinski, Michal	2019	NIM Marketing Intelligence Review
Who Shares Fake News in Online Social Networks?	Burbach, Laura e Halbach, Patrick e Ziefle, Martina e Calero Valdez, André	2019	Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization
Human and Computer Personality Prediction From Digital Footprints	Hinds, Joanne e Joinson, Adam	2019	Current Directions in Psychological Science
Mining Facebook Data for Personality Prediction: An Overview	Marengo, Davide e Settanni, Michele	2019	Digital Phenotyping and Mobile Sensing
Predicting Consumers' Decision-Making Styles by Analyzing Digital Footprints on Facebook	Chen, Yuh-Jen e Chen, Yuh-Min e Hsu, Yu-Jen e Wu, Jyun-Han	2019	International Journal of Information Technology & Decision Making
Analysis of factors that influence customers' willingness to leave big data digital footprints on social media: A systematic review of literature	Muhammad, Syed Sardar e Dey, Bidit Lal e Weerakkody, Vishanth	2018	Information Systems Frontiers
Predicting the Big 5 personality traits from digital footprints on social media: A meta-analysis	Azucar, Danny e Marengo, Davide e Settanni, Michele	2018	Personality and Individual Differences
Emerging trends in personality identification using online social networks—a literature survey	Kaushal, Vishal e Patwardhan, Manasi	2018	M Transactions on Knowledge Discovery from Data (TKDD)
User data privacy: Facebook, Cambridge Analytica, and privacy protection	Isaak, Jim e Hanna, Mina J	2018	d
What your facebook profile picture reveals about your personality	Segalin, Cristina e Celli, Fabio e Polonio, Luca e Kosinski, Michal e Stillwell, David e Sebe, Nicu e Cristani, Marco e Lepri, Bruno	2017	IEEE

Profiling players using real-world datasets: Clustering the data and correlating the results with the big-five personality traits	Halim, Zahid e Atif, Muhammad e Rashid, Ahmar e Edwin, Cedric A	2017	IEEE Transactions on Affective Computing
Psychological targeting as an effective approach to digital mass persuasion	Matz, Sandra C e Kosinski, Michal e Nave, Gideon e Stillwell, David J	2017	Proceedings of the national academy of sciences
Computer-based personality judgments are more accurate than those made by humans	Youyou, Wu e Kosinski, Michal e Stillwell, David	2015	Proceedings of the National Academy of Sciences
Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines	Kosinski, Michal e Matz, Sandra C, Gosling, Samuel D e Popov, Vesselin e Stillwell, David	2015	The American psychologist
Manifestations of user personality in website choice and behaviour on online social networks	Kosinski, Michal e Bachrach, Yoram e Kohli, Pushmeet e Stillwell, David e Graepel, Thore	2014	Machine learning (Springer)
Tracking the digital footprints of personality	Lambiotte, Renaud e Kosinski, Michal	2014	Proceedings of the IEEE
Recognising personality traits using Facebook status updates	Farnadi, Golnoosh e Zoghbi, Susana e Moens, Marie-Francine e De Cock, Martine	2013	Seventh International AAAI Conference on Weblogs and Social Media
Towards automated personality identification using speech acts	Appling, Darren Scott e Briscoe, Erica J e Hayes, Heather e Mappus, Rudolph L	2013	Seventh international AAAI conference on weblogs and social media
Predicting personality with social behavior	Adali, Sibel e Golbeck, Jennifer	2012	IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
Predicting personality with social media	Golbeck, Jennifer e Robles, Cristina e Turner, Karen	2011	CHI'11 extended abstracts on human factors in computing systems

5.1 Resultados

Nesta seção estão descritos os resultados obtidos a partir da leitura e análise dos trabalhos selecionados, a fim de responder às questões principais e secundárias de pesquisa.

5.1.1 (QP_1) Traços de personalidade e comportamentos de usuário em redes sociais

Com o desenvolvimento das redes sociais, uma variedade de abordagens foram desenvolvidas para entender a relação entre traços de personalidade e o comportamento humano em redes sociais. Uma parte desses estudos é baseada em dados explícitos do perfil, que geralmente são coletados durante o processo de inscrição em um serviço online. No entanto, a grande maioria dos trabalhos selecionados utilizam a base de dados *myPersonality*, com dados de milhões de usuários do Facebook coletados ainda em 2007 e com uma ampla variedade de dados relacionados à personalidade e o comportamento na rede social.

Por esse motivo, antes de explorar essa questão principal da pesquisa, é importante introduzir o principal conjunto de dados utilizados pelos estudos que envolvem redes sociais e traços de personalidade. O *myPersonality*, desenvolvido por Stillwell e Kosinski (2004), foi o primeiro projeto que fez uso de um conteúdo grande gerado pelas mídias sociais, com mais de 7,5 milhões de perfis de usuário do Facebook rotulados pelo modelo dos Cinco Grande Fatores. O resultado foi a criação de um dos maiores bancos de dados de pesquisa em ciências sociais da história. O aplicativo conectado ao Facebook realizava uma pesquisa psicológica, a partir de um questionário de personalidade. Em troca o aplicativo enviava *feedback* sobre as pontuações de cada usuário participante na pesquisa. O aplicativo ficou ativo até 2012. Além dos dados de traços de personalidade, o aplicativo também coletava alguns dados comportamentais a partir de pegadas digitais ativas, tais como curtidas, rede de amigos, atualizações de status e dados demográficos. Esses dados coletados foram anonimizados e as amostras foram compartilhadas com colaboradores acadêmicos registrados em todo o mundo, resultando em 57 publicações científicas em revistas especializadas (KOSINSKI, 2018a). Porém, em abril de 2018, os autores decidiram parar de compartilhar os dados com outros pesquisadores, pois, segundo eles, o projeto se tornou oneroso em termos de dedicação e adequação a várias regulamentações de privacidade. Como vamos revisar nessa subseção e nas próximas, diversos estudos foram realizados com sucesso nesse conjunto de dados. Um exemplo é um estudo subsequente de Kosinski et al. (2015b) que detectou que as informações de pegadas digitais dos usuários na forma de preferências (a partir de curtidas) são um bom preditor de personalidade. É indiscutível a contribuição desse projeto no estudo da psicologia, ao mesmo tempo que a revolução digital abriu novas perspectivas para estudiosos interessados em entender os seres humanos e melhorar a vida das pessoas. Entretanto, segundo Kosinski (2018b), o projeto *myPersonality* foi fechado em 2012 devido à falta de tempo dos autores para mantê-lo e foi relevante para a criação de novas leis de privacidade da União Europeia e dos Estados Unidos.

Utilizando esse conjunto de dados, o trabalho desenvolvido por Kosinski et al. (2014) examina como a personalidade se manifesta no comportamento online dos usuários pelos sites em que navegam e por suas atividades na rede social Facebook. Como a navegação na Internet de modo amplo é uma atividade percebida como privada pelos usuários (onde pegadas digitais são deixadas de forma passiva), segundo os autores, a atividade de consumir conteúdos digitais não sofre pressão social externa. Isso significa que o histórico de navegação de uma pessoa apresenta-se de maneira natural às preferências e personalidade de uma pessoa, ou seja, a atividade de navegar na Internet de forma ampla é um preditor potencialmente assertivo da personalidade de uma pessoa. Atributos como a frequência de uso, a distribuição de interações sociais públicas, a quantidade de fotos enviadas e a densidade da rede de amizade são dados menos prováveis de serem afetados por tentativas conscientes dos usuários de controlar sua imagem. Esse não é o caso das publicações de atualizações de *status* nas redes sociais, das imagens enviadas ou mesmo das interações sociais ativas (curtidas ou compartilhamentos públicos, por exemplo), que podem conter elementos de auto-aprimoramento social. Dessa forma, Kosinski et al. (2014) realizou dois estudos principais sobre a correlação entre traços de personalidade e comportamento de usuários em redes sociais:

1. **Personalidade e preferências de sites**: o objetivo deste estudo foi examinar como a personalidade de uma pessoa se reflete em seus hábitos de navegação na Internet. Um questionário foi elaborado para este estudo, solicitando a frequência com que 10.897 pessoas visitavam 23 sites previamente selecionados. Esses sites foram selecionados para serem potencialmente informativos sobre um personalidade do visitante. Para isso, esses sites não poderiam ser nem muito populares nem muito obscuros. Isso porque sites extremamente populares atraem visitantes de todos os tipos de personalidade e, portanto, não são informativos. Por outro lado, sites obscuros não atraem uma fração razoável de pessoas e, portanto, não são discriminatórios. Em comparação ao grupo que respondeu o questionário, um outro grupo de dados foi coletado diretamente de pegadas digitais deixadas por cerca de 153.000 usuários ao curtirem esses determinados sites no Facebook.

   Por exemplo, foi assumido (e posteriormente confirmado pelos resultados) que um site de letras de músicas seria atraente para pessoas sociáveis (altos níveis de extroversão). Ao comparar a assertividade entre o questionário e o comportamento coletado a partir de pegadas digitais (*Page Likes*) e as respostas ao questionário específico de preferência de navegação (WPQ), os autores concluíram que este site atrai um público que tende a ser liberal e artístico em vez de conservador e tradicional (ou seja, com alta abertura); espontâneo e flexível, em vez de bem organizado (ou seja, com baixa consciência); tímido e reservado em vez de extrovertido e ativo (ou seja, com baixa extroversão); e emocional em vez de calmo e relaxado (ou seja, com alto neuroticismo). No outro extremo da escala de abertura à experiência, o autor demonstra sites cuja população de usuários é estimada como mais conservadora e convencional e incluem sites de ofertas de produtos e sites de conteúdos sobre saúde, *fitness*, culinária e estilo.

A partir da análise de escala de correlação momento-produto de *Pearson* entre a personalidade agregada nos dois conjuntos de dados, os autores verificaram um alto nível de consistência nas amostras, fornecendo evidências que sustentam a validade dos seus estudos. Por exemplo, o valor de 0,83 foi obtido para o fator de conscienciosidade entre o conjunto de dados de *Page Likes* e o *WPQ*, indicando que esse fator está altamente correlacionado entre ambas as fontes de dados.

2. **Personalidade e dados demográficos de perfil**: esse estudo explora a relação entre a personalidade e os dados de perfil de um usuário de rede social, especificamente usuários do Facebook. Os autores também empregaram técnicas de regressão para prever as personalidades dos usuários com base nos perfis online. Segundo eles, apesar de as pessoas julgarem as personalidades de outras pessoas com base em seus perfis do Facebook ou histórico de navegação na Web, é possível que algumas características de personalidade sejam ignoradas ou mal interpretadas. Isso porque, como seres humanos, somos propensos a preconceitos que podem afetar a precisão de nossos julgamentos. Para entender a relação de personalidade e os dados demográficos de um perfil de rede social, foram obtidos dados de mais de 354.000 usuários do Facebook nos EUA, que utilizavam o Facebook há pelo menos 24 meses antes da coleta dos dados. O número total de amigos, eventos, atualizações de status, fotos, *tags* de fotos, associações a grupos e a densidade da rede de amizade foram as principais características observadas. Ainda que nesse trabalho relacionado os autores apontam que muitos usuários do Facebook tinham informações de perfil incompletas ou suas configurações de privacidade não permitiam acessar algumas partes do perfil, os autores encontraram relevantes correlações entre os dados demográficos de um usuário e os seus traços de personalidade. Por exemplo, utilizando o coeficiente de correlação de *Spearman*, foi percebido que os indivíduos liberais e abertos à experiência tendem a gostar de mais itens no Facebook (Figura 4, postar mais atualizações de status e participar de mais grupos, o que é consistente com a definição desse traço de personalidade. Usuários com o maior fator de neuroticismo também demonstraram alta correlação com a quantidade de interações sociais publicadas na rede. O mesmo não pode ser afirmado para pessoas com alto fator de contenciosidade, pois apresentam correlação inversa ao número de interações sociais publicadas (Figura 5. Essa situação corrobora com os estudos de traços de personalidade, já que são mais meticulosas nas suas ações.

Os dois estudos realizados pelos autores demonstram resultados positivos na correlação entre o comportamento de usuários de redes sociais e os seus traços de personalidades.

Burbach et al. (2019) buscaram entender a correlação entre personalidade e a propagação de *fake news* a partir do desenvolvimento de um modelo baseado em agente. Embora os autores tenham confirmado resultados de estudos precedentes (descritos nos parágrafos anteriores), onde pessoas com tendências narcisistas tenham mais amigos nas redes sociais e que a presença de maior consciência inibe as interações sociais em conteúdos, a aplicação do agente desenvolvido

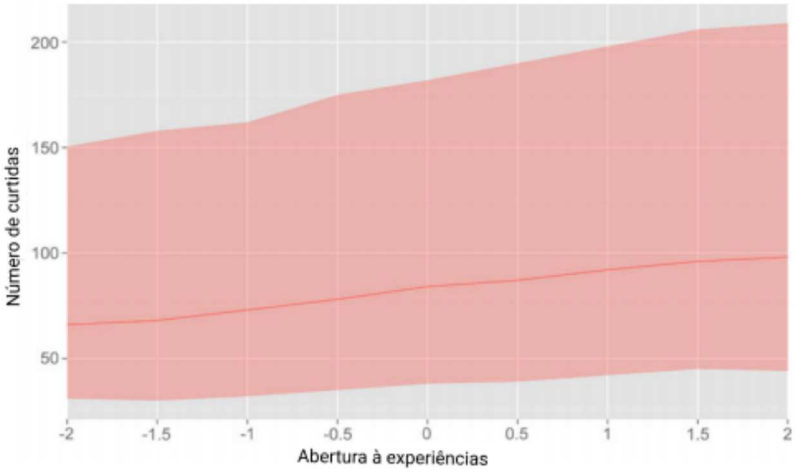

Figura 4: Número médio de curtidas de usuários caracterizados por diferentes níveis de abertura (traço de personalidade). O eixo y representa o intervalo interquartil médio do número de curtidas dos usuários no estudo de Kosinski et al. (2014).

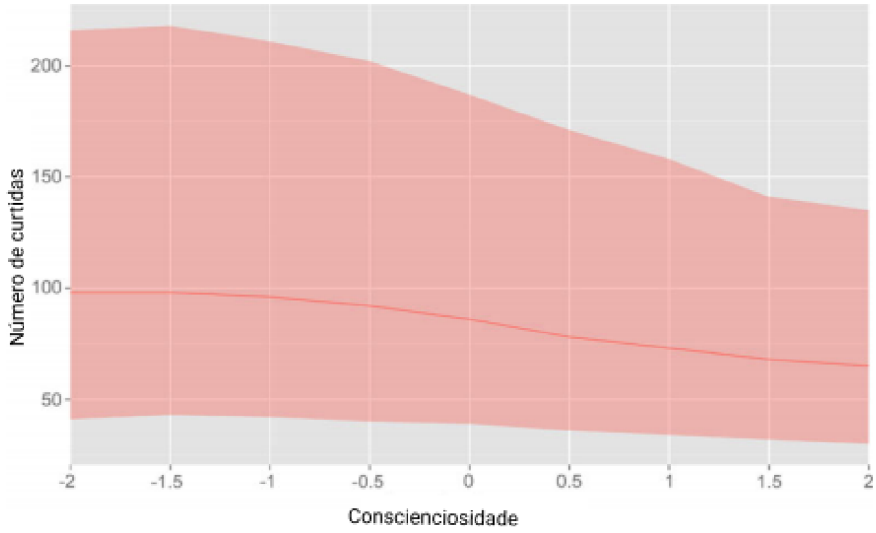

Figura 5: Número médio de curtidas para usuários caracterizados por diferentes níveis de conscienciosidade. O eixo y representa o intervalo interquartil médio do número de curtidas dos usuários no estudo de Kosinski et al. (2014)

não identificou um traço de personalidade específico na propagação das *fake news*. Inclusive, a densidade da rede foi identificada como mais relevante para a disseminação do que as diferenças de personalidade e comportamento dos indivíduos. Porém, deve-se ressaltar que os autores reconhecem uma série de limitações no estudo, seja pelas limitações intrínsecas da modelagem de um agente não humano, ou mesmo pela amostragem de personalidades de usuários coletada não possuir nenhum usuário com características extremas (por exemplo, um influenciador altamente narcisista).

5.1.2 (QP_2) A relação entre pegadas digitais em redes sociais e a personalidade de usuários

A interação dos usuários com as redes sociais produz uma série de pegadas digitais, como é revisado na Seção 3, consistindo em uma coleção de registros de atividades e dados textuais e visuais compartilhados. Esses registros são extensivamente coletados e extraídos para fins comerciais e representam uma fonte de dados preciosa para pesquisadores. Estudos recentes demonstraram que características obtidas com esses dados mostram conexões significativas com dados demográficos, comportamentais e características psicossociais dos usuários produtores de suas próprias pegadas digitais.

Lambiotte e Kosinski (2014) revisam a literatura acadêmica sobre o que é definido como *Big Social Data*, mostrando como essa coleção de registros difundidos de pegadas digitais podem ser usados para inferir personalidade. Ou seja, um dos principais *insights* oferecidos pelas pegadas digitais em redes sociais refere-se à previsibilidade dos traços psicológicos individuais. São as pegadas digitais deixadas nas redes sociais que possibilitam pesquisas em larga escala que visam entender a relação entre personalidade e comportamento humano. No artigo citado anteriormente, os autores sugerem que a personalidade não é manifestada apenas no offline, mas também no ambiente online e, portanto, as pegadas digitais podem ser usadas para detectar traços de personalidade. No estudo eles destacaram que o perfil do Facebook de um usuário não é puramente demográfico, como também contém registros robustos de pegadas digitais, principalmente nas manifestadas pela ação de "curtir"um conteúdo. Na Figura 6 pode ser observada a relação do traço de personalidade de abertura à experiências com pegadas digitais relacionadas a gostos musicais. Além disso, a proliferação de dispositivos móveis carregados com sensores significa que as atividades humanas offline também são cada vez mais rastreáveis. Por exemplo, estados como corrida ou caminhada podem ser deduzidos de dados do acelerômetro do dispositivo; a geolocalização pode ser estabelecida usando *WiFi* e GPS (Sistema de Posicionamento Global); e as interações sociais podem ser medidas por registros de mensagens de texto e telefonema. Dessa forma, é plausível perceber um potencial na combinação das pegadas digitais online e offline para inferência de personalidade. Os autores do estudo documentam que indicadores de mobilidade, como a distância percorrida, correlacionam-se significativamente com neuroticismo, enquanto indicadores da vida social, como o tamanho da rede social, correlacionada com a extroversão, em concordância com os resultados revisados na Seção 5.1.1.

O volume e a variedade de dados deixados por usuários de Internet em uma velocidade crescente foi discutido extensivamente em trabalhos acadêmicos. Porém, pouca consideração ainda é dada à dicotomia entre as motivações pessoais das pessoas geradoras e os debates profusamente debatidos sobre questões de privacidade e segurança. A revisão sistemática realizada por Muhammad, Dey e Weerakkody (2018) de 506 artigos revisados por pares, discute essa questão e revela que o comportamento pessoal (disposições psicológicas intrínsecas), fatores tecnológicos (vantagem e conveniência relativas), influência social (interação social, vínculos sociais

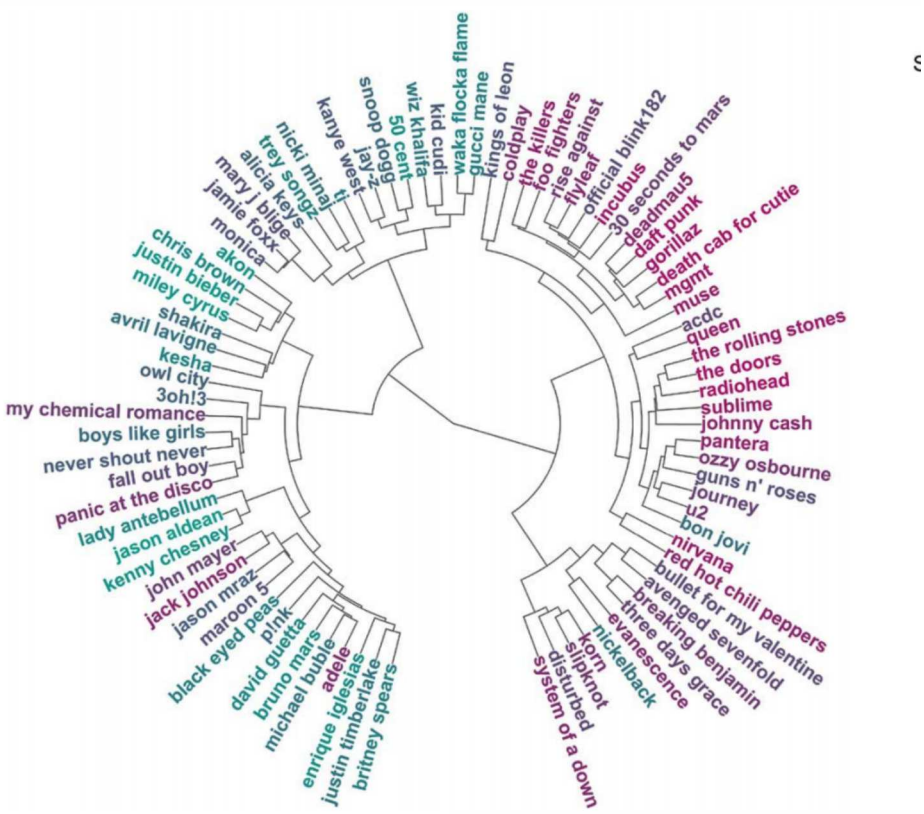

Figura 6: Dendrograma ilustrando a estrutura dos gostos musicais e a relação com o traço de personalidade da abertura. A estrutura foi produzida por Lambiotte e Kosinski (2014), usando um *cluster* hierárquico com as curtidas mais populares do Facebook na categoria musical. A escala de cores representa o traço de personalidade de abertura à experiência média, variando de conservador (azul ciano) para liberal (magenta)

e apoio social) e privacidade e segurança (risco, controle e confiança) são os principais fatores que influenciam as pessoas a deixarem suas pegadas digitais nas mídias sociais (Figura 7). Ao analisar-se as contribuições realizadas pelos autores nas disposições psicológicas de cada pessoa, revela-se dois fatores-chave de comportamento pessoal de auto-aprimoramento (autoeficácia e autoestima) e benefícios percebidos do prazer experiencial e sensorial que satisfazem necessidades dos usuários.

A partir da convergência entre ciências sociais e da computação e o desenvolvimento de métodos automatizados para extrair e relacionar as pegadas digitais com traços de personalidade, Azucar, Marengo e Settanni (2018) realizaram uma meta-análise para determinar o impacto de diferentes tipos de pegadas digitais na detecção de traços de personalidade. Embora existam vários modelos para descrever a personalidade, o estudo utilizou um dos mais bem pesquisados, bem vistos e com estruturas teóricas amplamente aceitas da personalidade, que é o Modelo dos Cinco Grande Fatores (ou Big 5) – revisado na Seção 2.2. Para estabelecer a magnitude da associação entre pegadas digitais e cada um dos cinco grandes traços de personalidade, o estudo realizou cinco meta-análises separadas, analisando 16 tamanhos de efeito para cada traço de personalidade. As correlações de tamanho médio de efeito variaram de 0,29 (agradabilidade)

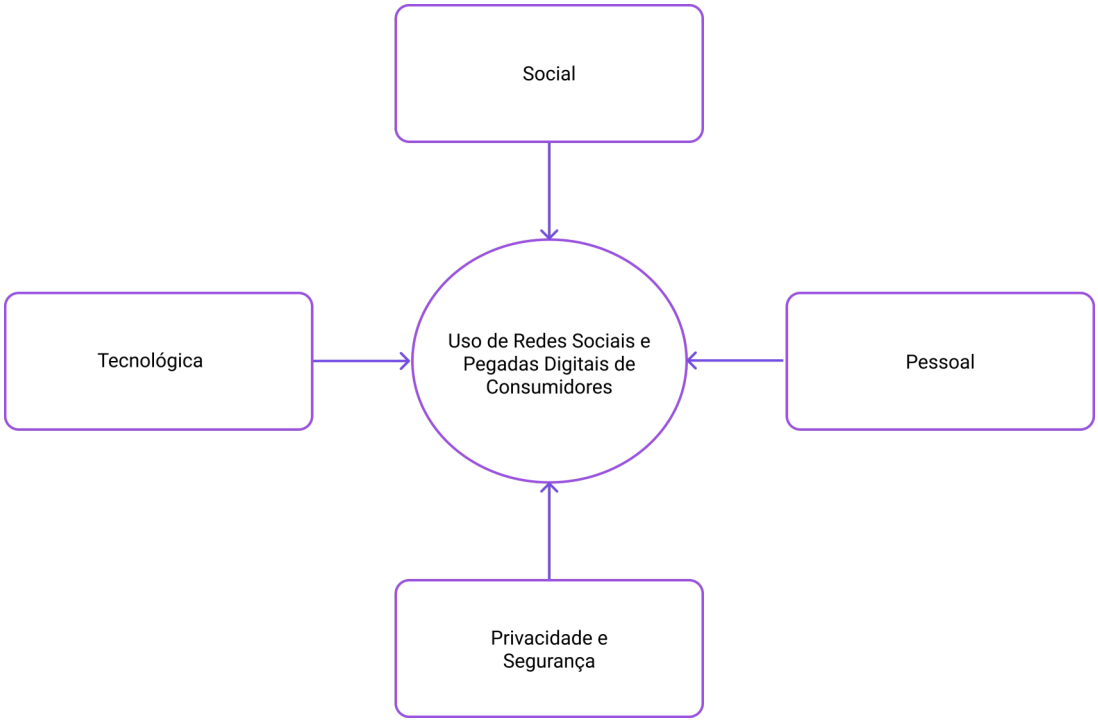

Figura 7: Modelo de classificação das motivações de deixar pegadas digitais desenvolvido por Muhammad, Dey e Weerakkody (2018)

a 0,40 (extroversão). No geral, o estudo indica que a precisão das detecção da personalidade é consistente entre os cinco principais traços e que a correlação entre pegadas digitais em redes sociais e traços de personalidade melhora a medida que as análises incluem maior dimensionalidade de pegadas digitais, como é detalhado na Seção 5.1.5.

Com base nesses estudos, é plausível perceber que as pessoas gastam um esforço considerável gerenciando as impressões que desejam deixar aos outros. Há também que se considerar o potencial de exploração dos diversos tipos de pegadas digitais que podem ser relacionadas ao julgamento da personalidade de usuários de redes sociais, não se limitando apenas as interações social clássicas e contínuas deixadas por elas. Para demonstrar isso, em uma especialização dos estudos relacionados à análise de personalidade através das pegadas digitais em mídias sociais, Segalin et al. (2017) estruturaram uma pesquisa a apenas um atributo visual deixado pelos usuários na rede social Facebook: a foto dos seus perfis. Nas suas descobertas, perceberam que pessoas extrovertidas e pessoas agradáveis tendem a publicar fotos com cores mais quentes e exibir muitos rostos em seus retratos, espelhando suas tendências a socializar (como pode ser visualizado na Figura 8); enquanto os neuróticos têm um prevalência de fotos em locais fechados.

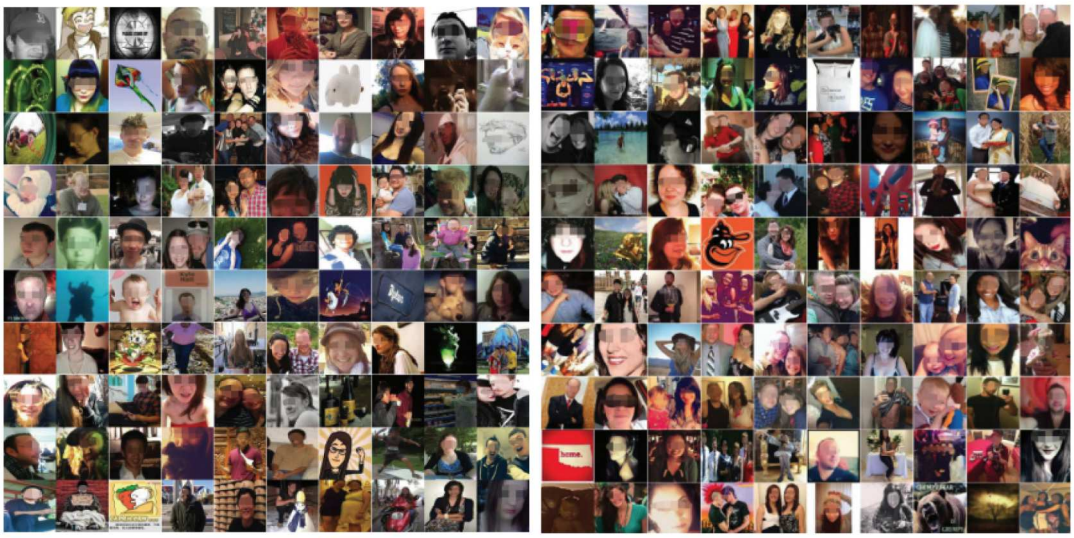

Figura 8: Na esquerda observa-se fotos de perfis de pessoas com fatores altos de extroversão, ao contrário do conjunto de fotos da direita que possuem baixa extroversão. As imagens foram disponibilizadas no estudo de Segalin et al. (2017) e o embaçamento foi utilizado por questões de privacidade.

5.1.3 (QP_3) Métodos e técnicas de mineração de dados e aprendizado de máquina empregadas para detectar automaticamente a personalidade

Reconhecer a personalidade de um indivíduo de forma eficiente e com alta precisão é um objetivo relevante para diversas disciplinas. As abordagens tradicionais para detectar a personalidade são entrevistas conduzidas por psicólogos e inventários ou questionários de autor-relato. Tais métodos, no entanto, são caros, propensos a erros e trabalhosos, ao modo que o avanço em áreas da ciência da computação, como processamento de imagem e inteligência artificial, introduziram técnicas automatizadas e tornaram o estudo em larga-escala uma realidade (KAUSHAL; PATWARDHAN, 2018).

O estudo de Marengo e Settanni (2019) compartilha um modelo de pesquisa comum nesse campo de estudo, consistindo de quatro etapas: (1) Pegadas digitais do usuário são coletadas e extraídas para a extração automatizada de *features*; (2) Informações sobre as características de personalidade de cada usuário (rótulos) são coletadas por meio de diferentes abordagens (por exemplo, pesquisa online, avaliação ecológica momentânea via aplicativos móveis); (3) Conjuntos de dados que combinam as *features* extraídas de pegadas digitais e informações dos usuários são exploradas para encontrar associações e treinar modelos destinados a prever os traços de personalidade, normalmente usando uma abordagem de aprendizado de máquina; e (4) modelos treinados concorrentes são comparados com base em sua precisão na detecção dos traços de personalidade dos usuários em novos conjuntos de dados independentes, levando à identificação do modelo com melhor desempenho.

Por sua vez, a meta-análise de Azucar, Marengo e Settanni (2018) examinou a literatura

com foco em estudos apresentando associações entre pegadas digitais na rede social Facebook e características do modelo dos Cinco Grandes Fatores. Na revisão, com base em 16 estudos independentes, a extroversão parece estar associada com a maior precisão de previsão geral (r = 0,40), seguida de abertura à experiências (r = 0,39), conscienciosidade (r = 0,35), neuroticismo (r = 0,33) e agradabilidade (r = 0,29). No geral, a precisão das previsões de personalidade baseadas na mineração de pegadas digitais extraídas do Facebook parecem ser moderadas. Os autores concluem que é previsível que o desenvolvimento de novas abordagens de coleta, integração, e análises (por exemplo, usando algoritmos de aprendizado profundo) contribuirão para tornar as previsões mais precisas e confiáveis.

De forma geral, existem muitas oportunidades para algoritmos de alta eficácia no julgamento da personalidade humana. O trabalho de Youyou, Kosinski e Stillwell (2015) ocupa um marco importante no estudo de potenciais melhorias nas interações sociais humano-computador. O estudo demonstra que os modelos baseados em computador são significativamente mais precisos que os humanos em uma tarefa cognitivo-social básica: julgamento da personalidade. E ao comparar a precisão dos julgamentos de personalidade entre humanos e algoritmos de aprendizado de máquina, utilizando uma amostra de 86.220 voluntários que responderam a um questionário de 100 itens sobre personalidade, os autores têm como resultado: (i) as previsões de computador com base em uma pegada digital genérica (curtidas no Facebook) são mais precisas (r = 0,56) do que aquelas feitas pelos amigos dos participantes no Facebook usando um questionário de personalidade (r = 0,49); (ii) os modelos de computador mostram maior concordância de avaliação em relação aos avaliados; e (iii) os julgamentos da personalidade do computador têm maior validade externa ao prever os resultados da vida, como uso de substâncias, atitudes políticas e saúde física; para alguns resultados, eles até superam os índices de personalidade auto-classificados. Essas descobertas destacam que as personalidades das pessoas podem ser previstas automaticamente e sem envolver habilidades sócio-cognitivas humanas. Além disso, a pesquisa deixou um espaço aberto onde os modelos de computadores poderiam ter um desempenho ainda mais decisivo nos seres humanos com a sofisticação dos modelos de computadores e a quantidade de pegadas digitais coletadas.

Como citado anteriormente, na revisão literária de Kaushal e Patwardhan (2018), a totalidade dos trabalhos revisados usaram essencialmente abordagem supervisionada para identificação de personalidade. Kaushal e Patwardhan (2018) ressalta que um dos principais problemas, no entanto, com a abordagem supervisionada, é a limitação de dados previamente anotados. Dados rotulados com tipos de personalidade são geralmente caros e demorados para coletar, pois implica na distribuição manual de questionários longos de aferimento de personalidade de acordo com a modelagem desejada. De qualquer forma, é importante ressaltar que a maioria dos estudos sobre identificação de personalidade usando perfis de redes sociais online considerou o *Big Five* como modelo padrão de fato para modelagem de personalidade - o mesmo utilizado nessa pesquisa. Os autores da revisão destacaram como desafio para ser enfrentado nessa área de pesquisa a ausência de conjuntos de dados mais bem definidos e publicados, que

poderiam ser usados em escala por pesquisadores e forneceriam uma base sólida para comparar resultados. Os autores concluem que a identificação de personalidade usando a análise de redes sociais é um domínio relativamente novo na pesquisa de aprendizado de máquina.

5.1.4 (QP_4) Métodos e técnicas de mineração de dados e aprendizado de máquina empregadas para agrupar usuários de redes sociais de acordo com seu comportamento e personalidade

Embora, e de forma ampla, seja observado uma amostragem pequena de trabalhos relacionados ao agrupamento de usuários de redes sociais por traços de personalidade – isso é observado na seleção de artigos relacionados para essa pesquisa, documentada na Seção 5 – agrupar o perfil das pessoas, o comportamento e a avaliação de suas personalidades atrai o interesse de uma variedade de campos de estudo. Essas iniciativas de trabalhos relacionados utilizam-se de diversas técnicas para direta ou indiretamente entender o comportamento das pessoas em diversos ambientes, e como suas características de comportamento e perfil correlacionam-se com seus traços de personalidade.

Halim et al. (2017) estudaram as ações de usuários de jogos individuais e como isso está correlacionado com a personalidade de cada jogador na vida real. Os resultados sugeriram que os participantes com alto índice de neuroticismo, extroversão e abertura à experiência eram agrupados em jogos com temática como tecnologia, forças armadas e sociedade. Ao combinar os resultados de personalidade com a demografia dos jogadores, verificou-se que os indivíduos que jogam jogos de computador desde a infância possuem maior índice de neuroticismo e abertura à experiência. Embora esse trabalho relacionado não esteja diretamente conectado à temática de redes sociais, o método utilizado é relevante para essa pesquisa. Por exemplo, para estudar a importância das diversas características presentes no conjunto de dados na formação de melhores agrupamentos, eles experimentam técnicas como a *Random Subset Feature Selection* (RSFS), *Principal Component Analysis* (PCA) e *Mutual Information* (MI). Em seguida, para criação dos agrupamentos, o estudo experimentou quatro técnicas de *clustering* - uma abordagem para análise estatística e estudada no capítulo 4, que agrupa os dados usando uma função de distância. Isso resultou em quatro formações de agrupamentos para cada conjunto de dados, com doze resultados de agrupamentos no total. Esse método permitiu ao estudo generalizar os resultados. Além disso, também foi realizado um estudo de agrupamento utilizando as características mais e menos significativas. As quatro técnicas de algoritmos para agrupamento podem resultar em formações de segmentações variáveis. Dessa forma, o algoritmo de maior eficácia na pesquisa foi o *Hierarchical Clustering*. Ele foi escolhido a partir da mensuração de índices de mensuração como o *Davies–Bouldin Index* (DBI) e *Silhouette Coefficient* (SC). No entanto, para estudar as propriedades de cada agrupamento (*cluster*), é necessária uma análise descritiva. A análise descritiva foi baseada no conjunto de características usado para gerar os agrupamentos. Dessa forma o estudo agrupou as características coletadas a partir do teste de personalidade aplicado

(no caso, o IPIP-NEO-120 - questionário de 120 itens) e os recursos do conjunto de dados dos jogos. A partir da análise descritiva, a pesquisa prosseguiu no entendimento de características correlacionados dos jogadores e sua personalidade, e no desenvolvimento de classificadores para detecção da personalidade.

Como contraponto a essa questão discutida, a maior parte dos trabalhos relacionados de análise e agrupamento de personalidade estão concentrados na língua inglesa, onde em sua maioria, utiliza técnicas de inferência de traços de personalidade a partir de um conjunto de dados rotulado, como são os casos das pesquisas de Kosinski, Stillwell e Graepel (2013), Wald, Khoshgoftaar e Sumner (2012), Adali e Golbeck (2012), Golbeck, Robles e Turner (2011), Golbeck et al. (2011), Ortigosa, Carro e Quiroga (2014), Markovikj et al. (2013) e outros detalhados na Seção 5.1.5. Dessa forma, pesquisas relevantes relacionadas a essa temática ainda são poucas. É o caso do trabalho relacionado de Tutaysalgir, Karagoz e Toroslu (2019), onde os autores agrupam conteúdos da rede social Twitter na língua turca, utilizando de um método de vetorização conhecido como *TF-IDF* (abreviação do inglês *Term Frequency–Inverse Document Frequency*, que significa Frequência do Termo–Inverso da Frequência nos Documentos) para extrair o tópico de um documento. Para obter resultados mais precisos, eles adicionam o modelo *Word2Vec*, que resolve problemas relacionados a representação de parágrafos que têm diferentes comprimentos e contextos. Nesse trabalho, os autores utilizaram-se de uma base rotulada de dados a partir de questionários, com um volume de 40 participantes. Para categorização dos traços de personalidade (cada dimensão), os autores discretizaram cada pontuação em quatro blocos: 0-25%, 25-50%, 50-75% e 75-100% - a mesma estratégia adotada na nossa pesquisa. O artigo, mesmo com uma amostra de dados pequena para validação, alcançou taxas de erro abaixo de 10% nas pontuações para "Abertura à Experiência", "Agradabilidade"e "Contenciosidade".

5.1.5 (QP_5) Dimensionalidade de dados utilizados nas pesquisas de agrupamento e inferência de personalidade de usuários em redes sociais

A personalidade é uma combinação de todos os atributos – comportamentais, temperamentais, emocionais e mentais – que caracterizam um indivíduo único (KAUSHAL; PATWARDHAN, 2018). Por essa razão, no campo de pesquisa da Computação da Personalidade, a modelagem da personalidade é uma tarefa que envolve uma heterogeneidade de dados. Entretanto, quais são os dados disponíveis para a comunidade acadêmica avançar no estudo de traços de personalidades de usuários de redes sociais? Essa seção destina-se a revisar qual o nível de acesso às pegadas digitais, ativas ou passivas, utilizadas por trabalhos relacionados.

Sabe-se que as redes sociais, em particular, servem como uma rica fonte de texto, bem como conteúdo não textual publicado pelos usuários (CHIN; WRIGHT, 2014). O *survey* de Kaushal e Patwardhan (2018), analisa as tendências emergentes na identificação de personalidade utilizando-se de redes sociais online. Os autores definem o conteúdo textual como uma

das principais fontes de identificação da personalidade de um usuário. Isso porque uma suposição central da psicologia da linguagem é que as palavras que as pessoas usam refletem seus traços de personalidade, ao mesmo tempo que é um conteúdo produzido em larga escala pelos usuários de redes sociais. Consequentemente, os pesquisadores fizeram uso extensivo de conteúdo textual disponível nos perfis de redes sociais para mineração de dados para estudos de personalidade. Dessa forma, a atualização de status é a dimensão mais presente em estudos de personalidade me redes sociais, como pode ser vista em Farnadi et al. (2013), Golbeck et al. (2011), Appling et al. (2013) e diversas outras pesquisas relacionadas.

Vários recursos de linguagem podem ser extraídos de atualizações de status e outros atributos textuais do perfil das redes sociais de seus usuários. Ainda na pesquisa relacionada de Kaushal e Patwardhan (2018), os autores citam os métodos mais utilizadas para extração de *features* dessa dimensão de dado específica, destacando-se a LIWC (em português Investigação Linguística e Contagem de Palavras), que produz dados estatísticos em 81 diferentes *features* agrupadas em cinco categorias como: a) contagens padrão (contagem de palavras, palavras com mais de seis letras, número de composições, etc.); b) processos psicológicos (processos emocionais, cognitivos, sensoriais e sociais); c) relatividade (palavras sobre tempo, passado, futuro); d) preocupações pessoais (ocupação, questões financeiras, saúde); e e) outras dimensões. Outros métodos citados são o a) *MRC Features*, calculado usando o banco de dados psicolinguístico do Conselho de Pesquisa Médica (COLTHEART, 1981), que consiste em mais de 150.000 palavras e suas características linguísticas e psicolinguísticas; b) *Speech acts*, a unidade básica de comunicação linguística humana (assertivas, comissivas, declarativas, diretivas e expressivas); c) *Parts-of-Speech* (POS), inclui o número médio de palavras em categorias gramaticais específicas (advérbios, adjetivos, verbos e pronomes); d) *H4LvD*, onde as palavras são classificadas usando uma escala de nível de intensidade, que é uma combinação de diferentes categorias de valência (positivo vs. negativo, forte vs. fraco e ativo vs. passivo; e e) análise de sentimento, com recursos que indicam a força de sentimentos positivos ou negativos de uma atualização de status.

Embora o conteúdo textual, na forma de atualização de status, seja a informação mais presente e acessível nas redes sociais, como revisamos na Seção 5.1.2, as pegadas digitais não se limitam a gerar apenas esse tipo de conteúdo. Existem várias características estruturais e comportamentais que podem ser extraídas para entender suas correlações com a personalidade. Os autores de Kaushal e Patwardhan (2018), descrevem as principais dimensões não-linguísticas recorrentemente utilizadas, tais como a) estruturais, número de amigos, a densidade da rede, ou seja, qual a porcentagem de possíveis arestas existentes entre amigos, centralidade e medidas de agrupamento; b) comportamentais, como o tipo de rede pessoal, associações com grupos, grau de revelação de informações privadas, número de diferentes funcionalidades usadas, fotos exibidas nos perfis individuais, presença de auto-retrato e um conjunto de várias outras medidas comportamentais introduzidas por Adali e Golbeck (2012); c) temporais, que podem incluir frequência de contato, frequência de aceitação/rejeição de amizade, diferença de número de

amigos durante um período de tempo, frequência de atualizações de status por dia ou número de atualizações de status postadas entre determinados horários. Em adição à essas dimensões não é plausível destacar também todas as pegadas digitais passivas discutidas nas questões principais anteriores a essa e no capítulo 3, onde destaque-se a navegação entre conteúdos da rede social.

Na Tabela 2 pode ser observada uma comparação entre diversos trabalhos relacionados ao estudo de personalidade em redes sociais. A Tabela foi estruturada para destacar o tamanho dos estudos realizados, a rede social analisada e, principalmente, as dimensões de dados coletadas para extração de informações para detecção da personalidade. Comparando os estudos, identifica-se uma maior concentração de trabalhos entre os anos de 2011 e 2013, onde o acesso ao *dataset myPersonality* (STILLWELL; KOSINSKI, 2004) era amplamente disponível. Outra informação relevante é a concentração de estudos nas principais redes sociais comerciais, com ampla concentração de estudos também no Facebook – pelo mesmo motivo citado anteriormente, o uso do *dataset myPersonality*. Além da concentração de estudos na língua inglesa, duas outras análises complementares destacam-se: a) nenhum dos estudos utiliza-se de qualquer pegada digital passiva, potencialmente por restrições de privacidade, b) os inventários de personalidades são baseados em questionários de tamanho médio (normalmente 45 questões). Dessa forma, encontra-se em aberto o espaço para estudos em larga-escala, com questionários reduzidos e coleta de pegadas digitais tanto ativas quanto passivas.

Tabela 2: Tabela comparativa de dimensões utilizadas em trabalhos relacionados ao estudo de personalidade em redes sociais

Artigo	***Dataset***	**Dimensões**
Golbeck, Robles e Turner (2011). Primeiro trabalho sobre a previsão da personalidade usando perfis sociais	**Facebook**. 279 pessoas. *Dataset* próprio. Inventário de personalidade com 45 questões.	Dados estruturais, pessoais, de atividades e preferências.
Quercia et al. (2012). Primeiro trabalho no Twitter. Estudou o relacionamento entre a personalidade e a previsão de personalidade dos seus usuários	**Twitter**. 355 pessoas. *myPersonality*.	Dados estruturais, pessoais, de atividades, preferências, LIWC e estados internos do Facebook.

Chapsky (2011). Trabalho inicial apresentado em gerar um modelo probabilístico de personalidade, que usa representações de conexões das pessoas com outras pessoas, lugares, culturas e ideias.	**Facebook**. 615 pessoas. *Dataset* próprio. Inventário de personalidade com 60 questões (NEO)	Cidade natal e interesses de filme e música.
Golbeck et al. (2011). Prevendo a personalidade de informações disponíveis publicamente em perfis de rede social.	**Twitter**. 279 pessoas. *Dataset* próprio. Inventário de personalidade com 45 questões.	MRC combinado com LIWC e análise de sentimento das publicações.
Bai, Gao e Zhu (2012). Previsão de personalidade em rede social chinesa.	**RenRen**. 209 pessoas. Inventário de personalidade com 44 questões. *Dataset* próprio.	Conteúdo publicado pelo usuário na rede social analisado a partir de vários dicionários de emoções.
Adali e Golbeck (2012). Examina em qual precisão medidas comportamentais podem ser usadas para prever personalidade	**Twitter**. 79 pessoas. *Dataset* próprio. Inventário de personalidade com 45 questões.	Incluiu 31 dados demográficos e 80 atributos baseados em LIWC.
Bachrach et al. (2012). A partir de uma amostra grande demonstra que combinando sinais de diferentes características é possível prever personalidade com confiabilidade	**Facebook**. 180.000 pessoas. *myPersonality*.	Duas categorias amplas - aspectos do perfil que dependem exclusivamente de ações (por exemplo, número de fotos enviadas) e aspectos do perfil que dependem das ações de um usuário e seus amigos em conjunto (por exemplo, número de vezes que o usuário foi marcado em fotos.

Wald, Khoshgoftaar e Sumner (2012). Prever traços de personalidade usando apenas dados demográficos e atributos baseados em texto extraído de perfis	**Facebook**. 537 pessoas. *Dataset* próprio. Inventário de personalidade com 45 questões.	Duas categorias amplas - aspectos do perfil que dependem exclusivamente de ações (por exemplo, número de fotos enviadas) e aspectos do perfil que dependem das ações de um usuário e seus amigos em conjunto (por exemplo, número de vezes que o usuário foi marcado em fotos.
Kosinski, Stillwell e Graepel (2013). Mostrou que curtidas podem ser usadas para automaticamente prever com precisão uma gama de atributos sensivelmente pessoais	**Facebook**. 58.000 pessoas. *myPersonality*.	Curtidas na rede social.
Verhoeven, Daelemans e De Smedt (2013). Relatar uma prova de conceito usando *ensemble learning* como um maneira de resolver o problema de aquisição de dados	**Facebook**. *myPersonality*.	Usado 2.000 trigramas frequentes como dimensões iniciais.
Farnadi et al. (2013). Medida F ponderada proposta como medida de avaliação	**Facebook**. *myPersonality*.	Usado quatro diferentes conjuntos de características: lexical(LIWC), medidas de redes (sociais), *timestamps* de atualização de status e outras medidas como postagens por usuário, letras maiúsculas, palavras repetidas e outras.

Markovikj et al. (2013). Explora padrões de linguística ricos usando grande conjunto de características para predição de personalidade.	**Facebook**. *myPersonality*.	Usou um grande conjunto de atributos (725), incluindo dados sociais e demográficos, para recursos lexicais, *speech tags* AFINN, H4Lvd e outros.
Alam, Stepanov e Riccardi (2013). Pacote de palavras em sequência como abordagem para prever personalidade.	**Facebook**. *myPersonality*.	Uni-gramas como recursos a partir de publicações sociais.
Appling et al. (2013). Conteúdo em atualizações de status para prever personalidade.	**Facebook**. *myPersonality*.	Focado em recursos simples do ato de fala em texto, com atributos derivados de questões como "tem uma palavra de pergunta?"e "Utilizou *emoticons*?".
Alam, Stepanov e Riccardi (2013). Pacote de palavras em sequência como abordagem para prever personalidade.	**Facebook**. *myPersonality*.	Uni-gramas como recursos a partir de publicações sociais.
Ye, Du e Zhao (2017). Predição de traços de personalidade de usuários em redes sociais, considerando não apenas as inter-relações entre os traços de personalidade de uma pessoa, mas também as inter-relações com os traços de personalidade de seus amigos.	**Facebook**. 17.327 pessoas. *myPersonality* e *Friends and Family dataset*.	Dados relacionados ao perfil, a estrutura de rede com os amigos e atualizações de status.

Wei et al. (2017). Predição da personalidade de usuários a partir de informações heterogêneas (incluindo o uso de idiomas próprios, avatar, emoticons e padrões responsivos).	**Twitter**. 3.162 pessoas. Dataset próprio. Inventário de personalidade com 44 questões.	Análise linguística a partir de atualizações de status.
Silva e Paraboni (2018). Investiga métodos recentes para representação de texto como uma alternativa possível ao reconhecimento de personalidade padrão baseado no conhecimento psicolinguístico dependente da linguagem. Estudo aplicada em língua portuguesa.	**Facebook**. 1.013 pessoas. Dataset próprio. Inventário de personalidade com 44 questões (IGFP-5).	Análise linguística a partir de atualizações de status.
Bhavya, Pillai e Guazzaroni (2020). Estudo que utilizada de algoritmos de *Deep Learning* para detecção de personalidade através de textos publicados em redes sociais online.	**Facebook**. 115,864 pessoas. *myPersonality*.	Análise linguística a partir de atualizações de status.
Segalin et al. (2017). Proposta de abordagem de classificação para reconhecer automaticamente traços de personalidade utilizando recursos visuais.	**Facebook**. 11.736 pessoas. *myPersonality*.	Análise de recursos visuais como foto de perfil.
Tutaysalgir, Karagoz e Toroslu (2019). O estudo tem como objetivo detectar personalidades aplicando técnicas de mineração de dados e aprendizado de máquina em conteúdo de língua turca.	**Twitter**. 40 pessoas. Dataset própria. Não cita inventário utilizado.	Análise linguística a partir de atualizações de status.

5.1.6 (QP_6) A detecção da personalidade como ferramenta para influenciar o comportamento de usuários nas redes sociais

Devido à crescente popularidade das redes sociais online, os pesquisadores começaram a analisar a possibilidade de prever a personalidade de seus usuários a partir do conteúdo rico produzido por eles digitalmente. Conforme revisado na Seção 5.1.3, estudos recentes sugerem que as redes sociais online são, de fato, um meio relevante e válido de comunicar a personalidade. Mas quais são as oportunidades relevantes nesse campo de pesquisa? Essa seção revisa como a detecção da personalidade é uma potencial ferramenta para influenciar o comportamento de usuários, especificamente, no contexto das redes sociais.

Na pesquisa literária de Kaushal e Patwardhan (2018), que revisa o estado da arte na identificação de personalidade de usuários de redes sociais, os autores descrevem sobre os muitos aspectos que os traços de personalidade representam no comportamento do usuário: atitude em relação às máquinas, desempenho geral no trabalho, capacidade e motivação acadêmica, distúrbios psicológicos, preferências musicais, avaliação de agentes conversacionais, capacidade de liderança, capacidade de vendas, eficácia como professor, habilidades de marketing e assim por diante. Consequentemente, a identificação da personalidade continua a encontrar aplicações interessantes nessas áreas relacionadas. Alguns exemplos incluem sistemas de recomendação, sistemas de detecção de fraude, atribuição de autoria de conteúdos, sistemas de conversação, sistemas de treinamento, sites de namoro para prever relacionamentos bem-sucedidos e verificação da compatibilidade, avaliação de candidatos à empregos, projetos de interfaces de usuário personalizadas, análise de conversas de suspeitos de terrorismo, marketing direcionado, sistema de sugestões de amigos em redes sociais e assim por diante.

A aplicabilidade de todas essas oportunidades está relacionada ao fato de o que convence uma pessoa pode não necessariamente convencer outra. O trabalho relacionado de Matz e Kosinski (2019) revisa o estudo das personalidades como ferramenta de comunicação de persuasiva em massa e como utilizar esse mecanismos em grandes grupos de pessoas para elas acreditarem e agirem do ponto de vista do comunicador. No estudo, destaca-se o potencial da utilização dessa ferramenta pelos governos para incentivar comportamentos saudáveis pelos profissionais de marketing – para adquirir e reter consumidores – e pelos partidos políticos – para mobilizar os seus potenciais eleitores. Essa combinação de oportunidades só é possível porque a comunicação persuasiva é particularmente eficaz quando adaptada às características psicológicas únicas das pessoas e suas motivações.

Em 2019 o investimento total em publicidade digital no Facebook, alcançou 1,57 bilhão de dólares, o que representou um aumento considerável em relação ao montante do ano anterior de 1,1 bilhão, como pode ser visto na Figura 9. Enquanto o Facebook não disponibiliza como as pegadas digitais dos usuários e os potenciais traços de personalidades inferidos são utilizados em seus algoritmos e o quanto o estudo da psicologia em larga escala está relacionado ao sucesso do Facebook como empresa de publicidade, Matz e Kosinski (2019) desenvolveram

e contrataram três campanhas do Facebook que alcançaram mais de 3,5 milhões de pessoas. Nas campanhas publicadas, eles utilizaram traços de personalidade previstos a partir de curtidas realizadas pelos usuários segmentados nas campanhas para testar a eficácia da publicidade adaptada à personalidade delas. Como resultado, eles constataram correspondência significativa entre o conteúdo dos apelos persuasivos e a personalidade das pessoas, mensurada pelo comportamento dos usuários de acordo com taxas de cliques nos anúncios e em compras realizadas. Por exemplo, os anúncios persuasivos que correspondiam ao nível de extroversão ou abertura à experiência das pessoas resultaram em até 40% mais cliques e em até 50% mais compras quando comparado aos anúncios que não foram adaptados à personalidade dos usuários. Dessa forma, os autores destacam como as recomendações de marketing e produto poderiam ser mais relevantes ao adicionar dimensões psicológicas ao modelos atuais de comunicação. Como exemplo, eles citam que anúncios de seguros online podem enfatizar a segurança ao comunicar-se com usuários emocionalmente instáveis.

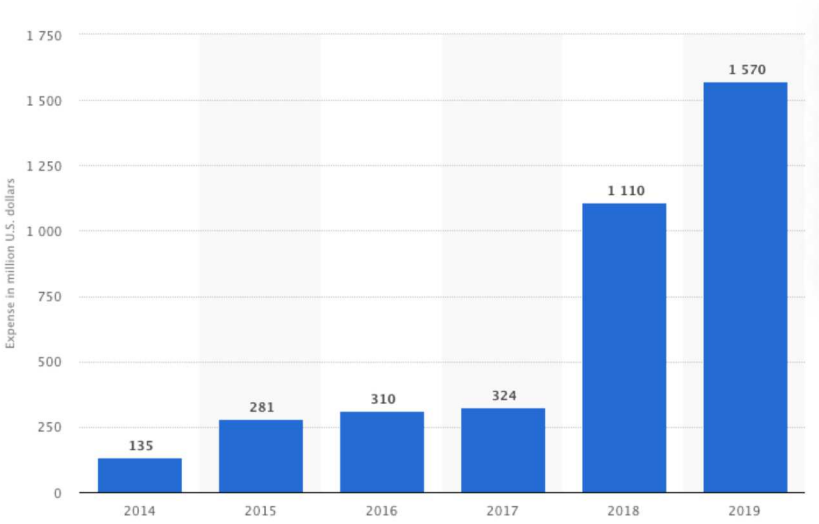

Figura 9: Investimento em publicidade digital no Facebook (STATISTA, ????)

Em um artigo anterior, Matz et al. (2017) detalham como esses experimentos foram realizados. Cada campanha de publicidade foi considerada um estudo diferente. O estudo 1 demonstra os efeitos da persuasão psicológica sobre comportamento de compra das pessoas. Eles adaptaram as mensagens de publicidade persuasivas para um varejista de beleza do Reino Unido para os destinatários com diferentes índices de extroversão. A Figura 10 mostra anúncios diferentes criados sob medida para cada um dos extremos. Em média os usuários que receberem publicidade congruente com esse traço de personalidade tiveram 1,54 vezes mais chances de realizar uma compra quando comparado aqueles em condições incongruentes. O estudo 2 replica e amplia as descobertas anteriores ao adaptar as mensagens de publicidades persuasivas de um aplicativo de palavras cruzadas para nível de abertura à experiência. Na Figura 11 é possível ver as diferenças nos anúncios criados, onde os usuários nas condições congruentes tiveram 1,38 vezes mais probabilidade de clicar no anúncio e 1,31 vezes mais chances de instalar o aplica-

tivo. No estudo 3, os autores combinaram essas descobertas, e utilizando-se de públicos-alvos pré-definidos para as campanhas, onde já se conhecia previamente que o traço de personalidade de introversão era o com maior índice de jogadores de jogos similares. Eles oferecem um jogo de quebra-cabeças com anúncio que tinha congruência com esse traço de personalidade, e compararam essa publicidade com o anúncio padrão já utilizado pelo jogo. Eles verificaram que a conversão foi de 1,2 vezes maior quando a mensagem publicitária persuasiva foi adaptada ao perfil psicológico do comportamento preexistente do público-alvo.

O estudo desenvolvido por Chen et al. (2019) oferece uma abordagem similar à anterior, prevendo estilos de tomada de decisão do consumidor analisando pegadas digitais no Facebook para ajudar empresas a reduzir custos de marketing e prover a satisfação do cliente. Eles oferecem um modelo para aplicabilidade de soluções relacionadas ao tema, a partir da execução das seguintes tarefas: (i) projetar um processo para prever estilos de tomada de decisão do consumidor com base na análise de pegadas digitais; (ii) desenvolver técnicas relacionados à previsão do estilo de tomada de decisão do consumidor; e (iii) implementar e avaliar um mecanismo de previsão de estilo de tomada de decisão do consumidor. Em um experimento prático, os autores utilizaram questionários e dados coletados de pegadas digitais (curtidas, status, fotos e vídeo) de 3304 participantes em 2018, 2644 dos quais foram selecionados aleatoriamente como treinamento do conjunto de dados, com os 660 participantes restantes formando um conjunto de dados de teste. Os resultados experimentais indicaram que a precisão aumentou para 75,88% e provaram que a abordagem proposta no estudo pode prever de maneira eficaz os estilos de tomada de decisão dos consumidores. Segundo os autores, essa abordagem não está restrita apenas à publicidade digital, podendo ser aplicada também a outros setores. Além disso, eles destacam que cada vez mais as pegadas digitais são apresentadas em foto e vídeo (sem texto), dimensões ainda pouco exploradas e com potencial de pesquisa, principalmente para entender a aplicabilidade de soluções similares a apresentadas por eles, com outras categorias de dados.

Essa revisão demonstra que os resultados obtidos pelos trabalhos relacionados anteriores fornecem evidências convergentes para a eficácia do direcionamento psicológico no contexto da persuasão em massa digital da vida real. Dessa foram, o esforço de adaptar produtos digitais, e especificamente redes sociais, para a detecção da personalidade, permite utilizá-la como ferramenta para influenciar o comportamento de usuários e suas escolhas reais.

5.1.7 (QS_1) Questões modernas relacionadas à privacidade no uso de pegadas digitais ativas e passivas para objetivos comerciais e de pesquisa

Como foi revisitado nas questões principais desta revisão sistemática, as conclusões de diversos estudos apresentam várias implicações importantes para o projeto de uma nova geração de sistemas, capazes de adquirir uma compreensão muito mais profunda dos usuários com base em seus traços pessoais e adaptar as interações com os usuários com base em sua personalidade. Dessa forma, como a personalidade distingue-se exclusivamente de um indivíduo para outro,

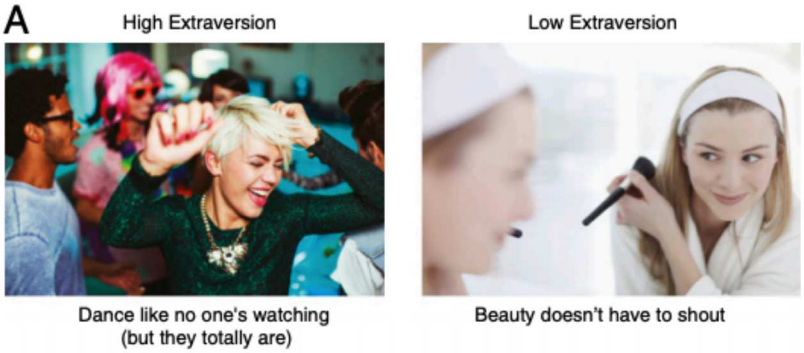

Figura 10: Exemplos de anúncios destinados a públicos caracterizados por alto e baixo índice de extroversão

Figura 11: Exemplos de anúncios destinados a públicos caracterizados por alto e baixo índice de abertura à experiência

há uma discussão ética em relação à coleta, manipulação aplicação desses dados sensíveis.

O *targeting* (ou segmentação), a partir de características psicológicas de um usuário de rede social, descreve a prática de extrair perfis psicológicos das pessoas a partir de suas pegadas digitais para influenciar suas atitudes, emoções ou comportamentos em escala e criar soluções em uma variedade de domínios, como pode ser observado na Figura 12. Essa prática ganhou destaque global durante a eleição presidencial de 2016 nos EUA, depois de uma empresa chamada Cambridge Analytica afirmar ter extraído perfis psicológicos de milhões de usuários do Facebook para atingir com publicidade psicologicamente adaptada a cada perfil de eleitor (ISAAK; HANNA, 2018).

A crescente quantidade de dados coletados a partir de inúmeros sensores (localização, saúde, voz) e serviços (Spotify, Twitter, Facebook, Google), que coletam pegadas digitais em tempo real, acabam criando registros extensos de nossos hábitos e preferências pessoais, sem a percepção dos seus usuários. Os autores Matz, Appel e Kosinski (2020), na sua pesquisa sobre o assunto, discutem a confusa relação entre dados públicos e privados, bem como a contínua utilização de práticas desatualizadas e sem consentimento prévio por parte dos usuários de produtos digitais. Segundo a revisão realizada por eles, a explosão de pegadas digitais (20 quintilhões de bits são criados todos os dias) inaugurou uma nova era para a privacidade. Além disso, uma

Figura 12: Como funciona a segmentação psicológica? Figura projetada a partir da pesquisa de Matz, Appel e Kosinski (2020)

vez que um conteúdo é publicado em uma rede social, torna-se quase impossível tornar essa informação privada novamente. O sistema de regulamento de privacidade, segundo os autores do artigo, estão limitados às estrategias de utilização de instrumentos de aviso e consentimento (também conhecido como consentimento informado). Nesse sentido, os autores apontam que essa estratégia é inadequada para proteger a privacidade das pessoas por vários motivos, como a) o cenário moderno de privacidade estar mais complexo do que nunca, apresentando políticas de privacidade que mudam frequentemente e envolvem vários terceiros com políticas separadas (exemplo um dado coletado por geolocalização de um aplicativo de transporte, pode ser utilizado para melhorar o algoritmo de recomendação de um aplicativo de rede social); b) a maioria das pessoas não tem conhecimentos para tomar decisões sobre privacidade que são do seu interesse (por exemplo, elas podem confundir dados que poderia revelar informações íntimas por inócuo porque elas não sabem sobre as etapas de processamento envolvidos e envolvidos) e porque são vítimas de vieses cognitivos bem estabelecidos que os levam a tomar decisões rápidas (por exemplo, avaliar instantaneamente o compartilhamento de informações a parir da ótica da alta gratificação em contrapartida aos riscos à privacidade a longo-prazo); c) combinações e usos futuros de dados são imprevisíveis na fase de consentimento (por exemplo, mesmo usuários experientes e racionais não podem entender o que eles consentem); d) Instrumentos de consentimento se concentram no consentimento individual, enquanto suas consequências também afetam os outros (por exemplo, indivíduos com consentimento podem ser usados para traçar indivíduos sem consentimento com características observáveis semelhantes; e) os mecanismos de consentimento mais atuais são binários e não possuem níveis adequados de granularidade (por exemplo, um o usuário geralmente precisa concordar com todas as práticas de dados ou descontinuar o uso de um serviço).

Escândalos de vazamento de dados sensíveis invariavelmente resultam em apelos a regulamentações rígidas e supervisão governamental. O GDPR (Regulamento Geral de Proteção de Dados), da União Europeia, está entre os mais estritos regulamentos de proteção em todo o mundo e o primeiro a mencionar o conceito de "criação de perfil"e seu uso na tomada de deci-

sões automatizadas (TIKKINEN-PIRI; ROHUNEN; MARKKULA, 2018). No Brasil, a LGPD (Lei Geral de Proteção de Dados), que ainda têm seu lançamento previsto para 2021, segue a mesma estratégia adotada pela GDPR. Ela também foi criada pela necessidade do país ter uma lei específica sobre proteção de dados pessoais - utilizados exaustivamente como moeda de sustentação de diversos modelos de negócios digitais (MALDONADO; BLUM; BORELLI, 2019). Ambos os regulamentos estão fortemente baseados no conceito de penalidades às empresas infratoras ao princípio da transparência, que exige que elas divulguem - em termos claros e simples - não apenas que tipo de dados são coletados, mas também para quais propósitos e se os mesmos estão sendo compartilhados com terceiros. Porém, corroborando com as motivações anteriores, o estudo de Obar e Oeldorf-Hirsch (2020) realizou uma investigação empírica do comportamento de leitura das políticas de privacidade disponibilizada por redes sociais. Os resultados revelaram que 74% do grupo de pessoas (N=543) ignoraram a leitura e acessaram diretamente o aplicativo. Aqueles que leram, dedicaram em média 73 segundos para tarefa.

Retornando à pesquisa de Matz, Appel e Kosinski (2020), os autores sugerem potenciais soluções para as questões modernas de privacidade. Uma delas é o desenvolvimento de um conceito onde a proteção à privacidade é colocada em um nível razoável por padrão, retirando do usuário o ônus de tomar decisões racionais e trabalhosas no momento de entrada em um produto digital. Outra sugestão é abordar segmentações de restrições de acesso aos dados. Por exemplo, um usuário poderia restringir o uso de seus dados em contextos específicos, como campanhas políticas. Além disso, os regulamentos deveriam ir além de proteger tipos de dados específicos, como dados de cartão de crédito e dados de saúde, em vez disso, e demonstrar que dados de compras podem estar intimamente vinculados a resultados de saúde. Dessa forma, esse tipo de dado mereceria melhor proteção do usuário. Segundo os autores, essas sugestões só têm efetividade se acompanhadas por um princípio focado em oportunidades e não em desafios: divulgação de dados sensíveis por escolha. Isso significa que ao colocar o interesse do usuário à frente do interesse comercial, criando-se uma relação de confiança, as tecnologias preditivas como a segmentação psicológica têm o potencial de melhorar a vida das pessoas. Dessa forma, a pesquisa questiona em quais contextos a segmentação psicológica é prejudicial e deve ser barrada e em quais contextos ela é benéfica para os usuários.

Na mesma linha, a pesquisa de Smith (2018) discute como os americanos percebem o uso de algoritmos modernos para tomadas de decisões importantes. No contexto das redes sociais, os algoritmos moldam e definem que conteúdo específico pode ser mais atraente para qualquer usuário, com base em conhecimento extraído de suas pegadas digitais. No âmbito dos estados afetivos, a pesquisa fornece ampla evidência de que os usuários de mídia social são regularmente expostos a conteúdo potencialmente problemático ou preocupante. Notavelmente, 71% dos usuários de mídia social dizem que sempre recebem conteúdo que os deixa com raiva - com 25% dizendo que veem esse tipo de conteúdo frequentemente. Da mesma forma, aproximadamente seis em cada dez usuários dizem que frequentemente encontram postagens que são excessivamente exageradas (58%). Quando é analisado o quão confortável os usuários se

sentem com plataformas de redes sociais utilizando os seus dados, os autores verificaram que 75% dos usuários acreditam ser aceitável que essas plataformas utilizam suas pegadas digitais para recomendar eventos que combinam com seus interesses pessoais. Por outro lado, 47% dos usuários analisados na pesquisa acreditam ser inaceitável utilizarem seus dados pessoais para receberem recomendações de publicidade sobre produtos ou serviços e apenas 35% deles pensam ser aceitável receber recomendações de campanhas políticas.

Há muitos desafios de privacidade a serem explorados e muitas hipóteses a serem desafiadas no âmbito de garantir segurança à privacidade de usuários de redes sociais. Além disso, existem poucas evidências sobre como pessoas com traços de personalidade distintos se comportam em relação às questões modernas relacionadas à privacidade no uso de pegadas digitais ativas e passivas para objetivos comerciais e de pesquisa.

5.1.8 Discussão dos Resultados

Os trabalhos analisados envolveram diferentes vieses de estudos de traços de personalidades e sua aplicabilidade em redes sociais. A revisão mostra, na Seção 5.1.1, como a personalidade de um usuário está intrinsecamente conectada as suas preferências e características de navegação em redes sociais. Essa conexão é potencialmente explorada por diversos negócios para aumentarem seus lucros, ao mesmo tempo que são uma fonte de dados preciosa para pesquisadores realizarem estudos da psicologia em larga-escala. Como foi discutido na Seção 5.1.2, isso é viável a partir da mineração de dados extraídos de pegadas digitais deixadas ativamente ou passivamente pelas pessoas ao utilizarem suas redes sociais online. Entre as técnicas e métodos mais utilizados para a produção de conhecimento e construção de previsibilidade de traços psicológicos, o aprendizado de máquina, utilizando-se de algoritmos supervisionados, foi validado na Seção 5.1.3 como aquele que possui maior concentração de pesquisas e trabalhos relacionados, alcançando resultados de julgamento de personalidade melhores do que o próprio julgamento humano. Embora essa área de pesquisa enfrente desafios pela ausência de datasets públicos para treinamento e validação de suas hipóteses. Como observado na Seção 5.1.4, os métodos e técnicas de aprendizado de máquina por agrupamento para o estudo de traços de personalidade ainda são exceções e com poucos artigos relevantes publicados, embora existam trabalhos relacionados em áreas similares, como o estudo do comportamento de usuários de jogos eletrônicos e seus traços de personalidade. Na Seção 5.1.5 foi percebido que os trabalhos relacionados estão limitados na falta de heterogeneidade dos dados coletadas, restringindo as pesquisas à pegadas digitas ativas, normalmente conteúdos de textos acessíveis publicamente. Dado todas essas condições e o estado-da-arte dessa área de conhecimento, conclui-se na Seção 5.1.6 que a detecção da personalidade é uma ferramenta que influencia o comportamento dos usuários nas redes sociais, embora muitos desafios e questões modernas de privacidade estejam em voga a partir de escândalos de vazamento de dados sensíveis que invariavelmente aumentam as restrições legais de acesso à essas informações (Seção 5.1.7).

O desafio na amostragem de dados é uma das maiores limitações que impedem os pesquisadores de dar uma contribuição ainda mais relevante nesse domínio. Isso cria a necessidade de disponibilizar à comunidade científica conjuntos de dados bem definidos, que poderiam ser usados por pesquisadores e forneceria uma base sólida para a comparação de resultados - o que se tornou ainda mais relevante desde a remoção do domínio público do principal *dataset* utilizado pelos pesquisadores até então. Além disso, os dados extraídos pelos pesquisadores são invariavelmente oriundos de pegadas digitais ativas, devido à falta de acesso aos dados percebidos como privados. Dessa forma, o uso de diferentes conjuntos de pegadas digitais, principalmente as passivas, ao serem extraídos e avaliados poderiam ter sua utilidade na identificação da personalidade de um usuário. Outro porém, refere-se aos questionários ou inventários de personalidade – que funcionam como instrumento de coleta de informações valiosas na construção de conjunto de dados rotulados, que ainda são normalmente baseados em questionários de tamanho médio. Essas questões anteriores não impedem a validação da detecção automatizada da personalidade das pessoas como uma potencial ferramenta para manipulá-las e influenciá-las a partir do desenvolvimento de aplicativos mais inteligentes e adaptáveis. Obviamente, além do desafio técnico, há questões éticas sobre o assunto a serem discutidas. A pesquisa sobre a utilização de aprendizado de máquina no estudo da personalidade nas redes sociais está em estágio inicial quando comparada ao crescimento anual das redes sociais digitais e as crescentes demandas de restrições de privacidade no acesso a dados por terceiros - que criam barreiras à pesquisa para acessar conjuntos de dados confiáveis.

Nessa revisão foi demonstrada que os dados das redes sociais digitais, especificamente perfis e curtidas de página, podem ser extraídos para criar classificadores capazes de detectar a personalidade do usuário com precisão razoável. No entanto, há potencialmente muitos dados sendo explorados pelas maiores empresas digitais sem o conhecimento acadêmico. Ao analisar às questões principais dessa pesquisa, percebe-se que o trabalho proposto nessa dissertação ao combinar a) dados coletados de pegadas digitais (ativas e passivas); b) dados demográficos; e c) dados de traços de personalidade a partir de questionários curtos de inferência. Ao agrupar usuários de redes sociais a partir da utilização de algoritmos aprendizado de máquina não-supervisionados – criando agrupamentos desconhecidos entre esses dados heterogêneos – é um dos primeiros a oferecer uma análise comparativa de padrões encontrados com tantas características de dados, por exemplo combinando variáveis de origens distintas, poder de investimento (demográfico), engajamento em conteúdos específicos da rede social (comportamento e pegadas digitais passivas) e extroversão (traços de personalidade). Entretanto, é válido ressaltar que este estudo apresenta limitações em relação às características específicas do produto e ao comportamento dos usuários dessa rede social específica utilizada na pesquisa. Além disso, há evidências de novas oportunidades para pesquisadores sociais, desde cientistas da computação (que podem desenvolver algoritmos para analisar massas de dados) até psicólogos (que podem testar teorias em novos domínios em grande escala), mas a comunicação interdisciplinar é rara nesse tópico de pesquisa (HINDS; JOINSON, 2019). Essa pesquisa reforça seus esforços para

fechar essa lacuna, criando uma parceria com pesquisadores experientes na área da psicologia.

No geral, em relação aos trabalhos relacionados, esta pesquisa contribui para os pesquisadores compreenderem como os traços de personalidade são expressos no comportamento humano dentro de redes sociais.

6 TRABALHO DESENVOLVIDO

Este capítulo descreve o desenvolvimento do trabalho proposto, que, especificamente, consiste em dois objetivos primários: (1) desenvolver agrupamentos criados a partir da intersecção de dados comportamentais (pegadas digitais ativas e passivas) e demográficos com os traços de personalidade de usuários de redes sociais, utilizando-se o modelo dos Cinco Grandes Fatores, inferidos a partir de um questionário reduzido de auto-relato, utilizando-se de algoritmos de aprendizado de máquina não supervisionados e (2) analisar qualitativamente e quantitativamente o processo de clusterização, verificando-se a criação de grupos significativos quando consideradas características socioafetivas no agrupamento, assim como pegadas digitais passivas, a fim de entender a qualidade dos grupos formados e o quanto eles fazem sentido.

Essa pesquisa também realiza a introdução de um novo conjunto de dados rotulados e com alta dimensionalidade referentes à combinação de dados de personalidade, dados comportamentais com características extraídas de pegadas digitais ativas e passivas, e dados demográficos de rede social na língua portuguesa. Intencionalmente esse estudo explora oportunidades em relação aos trabalhos relacionados (Seção 5), onde preferencialmente as pesquisas desenvolvidas em tópicos relacionados estão limitadas a baixa dimensionalidade de dados; a língua inglesa; a redes sociais específicas; a ausência do uso de pegadas digitais passivas; ao uso de questionários de inferência de personalidade extensos; e limitados a um escopo temporal onde os dados das principais redes sociais eram acessíveis para pesquisas de grande alcance 5.1.1.

Para o desenvolvimento do trabalho e alcance dos seus objetivos primários, um conjunto de dados de alta-dimensionalidade, contendo dados comportamentais e demográficos da rede social da Wedy[1] foi utilizado. O conjunto de dados inclui, além de pegadas digitais ativas, atributos referentes às pegadas digitais passivas. A Seção 6.4 apresenta com maiores detalhes as características dos dados disponibilizados pela rede social, como também a estratégia utilizada para coleta de dados de personalidade.

A hipótese de pesquisa deste trabalho sugere que os dados de personalidade dos usuários de rede social podem estar diretamente conectados com seu comportamento e suas informações demográficas, e que vão impactar na formação de grupos através da clusterização.

6.1 Etapas do Trabalho

Após as definições iniciais, a revisão bibliográfica, a solicitação de habilitação de estudos pelo Comitê de Ética, o desenvolvimento do sistema de aplicação do questionário de coleta de traços de personalidade, a revisão sistêmica de trabalhos relacionados e os estudos iniciais dos conjuntos disponibilizados para essa pesquisa, o trabalho seguiu com o desenvolvimento das atividades que estão descritas no cronograma na Tabela 14. Conforme este cronograma, após a

[1] A Wedy é uma startup de organização de casamentos que possui uma funcionalidade que conecta seus usuários em uma comunidade online. Maiores informações na Seção 6.3.2.

realização do seminário de qualificação, pretendia-se iniciar a coleta de dados de personalidade e prosseguir sequencialmente com as demais etapas. Entretanto, o cronograma proposto sofreu mudanças importantes e sequentes pelo cenário de pandemia que a sociedade encontrou-se no decorrer do desenvolvimento desse trabalho e a consequente baixa amostragem de dados passíveis de coletas devido à ausência total de eventos de médio porte como são os casamentos - matéria principal da rede social utilizada como objeto de estudo 6.3.2.

O cronograma condensa e divide a pesquisa em duas fases, como é demonstrado visualmente na Figura 13. Na primeira fase (F1), as seguintes etapas foram realizadas:

1. Aplicação da escala reduzida ER5FP (20 questões) para identificar e coletar a personalidade dos usuários da rede social Wedy. O ER5FP é uma medida de autorrelato breve e destinada a avaliar dimensões da personalidade baseada no modelo dos Cinco Grandes Fatores da Personalidade: abertura, conscienciosidade, extroversão, amabilidade e neuroticismo (LAROS et al., 2018). Os usuários acessaram o questionário de forma digital a partir da disponibilização de uma interface dentro de suas experiências no produto;

2. Combinação dos dados de personalidade com as informações que representam o comportamento dos usuários dessa rede social, representados por pegadas digitais (ativas e passivas) deixadas por eles;

3. Combinação dos dados de personalidade e comportamento com as informações demográficas cadastradas por esses usuários na rede social;

4. Utilização de algoritmos não-supervisionados de clusterização, bem como técnicas de engenharia de dados como limpeza e normalização de dados, extração de características e seleção de algoritmo a partir de técnicas de avaliação de modelos de aprendizado de máquina, para segmentar os grupos de acordos com as características pré-definidas nos passos anteriores;

Na segunda fase (F2), as seguintes atividades e métodos foram planejados e realizados:

1. Definição da quantidade de segmentações de usuários ideal a serem analisadas, de acordo com o conjunto de características selecionadas e a avaliação do desempenho de cada algoritmo de clusterização;

2. Extração das principais características presentes em cada grupo e análise descritiva das dimensões de dados presentes em cada segmentação em relação a cada traço de personalidade;

Figura 13: As duas fases do Trabalho Proposto

6.2 Viabilidade da Proposta

Como foi revisto nos trabalhos relacionados5, os dados comportamentais de usuários em redes sociais na forma de pegadas digitais são de difícil acesso e coleta. Isso se deve à ausência de dados rotulados disponibilizados publicamente, resultado das rígidas políticas de privacidade implantadas por órgãos regulamentadores (Seção 5.1.7) após os recentes escândalos de vazamento de dados sensíveis de usuários de redes sociais. Dadas essas limitações atuais, e enfatizando que esse estudo tem como o objetivo analisar o comportamento de usuários em redes sociais a partir de traços de personalidade, obter um conjunto de dados relevantes de acordo com políticas de seguranças de dados correntes foi uma prioridade da pesquisa. Dessa forma, a rede social Wedy, detalhada na Seção 6.3.2, foi escolhida para ser parceira nesse projeto, na qual os dados foram disponibilizados com consentimento de seus usuários a partir de restrições rígidas de políticas de privacidade. Os dados disponibilizados de forma anonimizada para essa pesquisa, contém os conjuntos de dados referenciados na Seção 6.4 no formato CSV (valores separados por vírgulas), de todos os usuários que responderam ao questionário de identificação dos traços de personalidade disponibilizados a partir de um aplicativo desenvolvido nessa pes-

quisa e disponibilizado dentro da rede social para realizar a coleta desses dados de personalidade (um exemplo está ilustrado na Figura 23).

6.3 Materiais

6.3.1 Instrumentos psicológicos

Um instrumento psicológico, no caso de testes objetivos, é essencialmente uma medida padronizada de uma amostra de comportamentos (ANDRADE, 2008). A confiabilidade de um instrumento psicológico pode ser verificada pelo campo de estudo psicométrico, que se preocupa com a teoria e a técnica da medição psicológica. O bem estabelecido e amplamente utilizado instrumento *Big Five Inventory* foi desenvolvido no final dos anos 80 por John, Donahue e Kentle (1991) com 44 itens de frases curtas, respondidos como autorrelato em cerca de 5 minutos e sendo capazes de avaliar os traços de personalidade com confiabilidade e validade.

Uma questão a ser considerada é a característica dos usuários da rede social estudada, onde o idioma praticado é o português do Brasil. Para essa questão, Andrade (2008) realizou a validação do *Big Five Inventory* para o contexto brasileiro, com um inventário reduzido para 32 itens, proposto como parte de um modelo de validação de instrumentos de personalidade, utilizando (1) evidência de validade fatorial, (2) diferenciação de escores nas variáveis sociodemográficas e (3) análise psicométrica de itens individuais.

No entanto, os pesquisadores geralmente são desafiados com restrições ambientais a usar escalas muito breves para validar suas suposições. No contexto de produtos digitais, especialmente dentro de redes sociais, os usuários têm o tempo como um recurso severamente limitado. Pela perspectiva dos usuários, o tempo que eles têm deve ser alocado para fazer o que eles desejam. Dessa forma, esta pesquisa atual está baseada nessa premissa do tempo e, consequentemente da atenção de um usuário de rede social, como recursos escassos. O instrumento selecionado para lidar com essa questão é a Escala Reduzida de Cinco Grandes Fatores de Personalidade (ER5FP), um questionário de autorrelato de curta duração, com apenas 20 itens (Apêndice A.1), destinado a avaliar dimensões da personalidade baseada no modelo dos Cinco Grandes Fatores da Personalidade: abertura, conscienciosidade, extroversão, amabilidade e neuroticismo. Os autores (LAROS et al., 2018) realizaram a validação dessa escala com 554 participantes entre 16 e 69 anos (M = 30,6, DP = 8,6). O modelo de medição de cada instrumento foi testado usando análise fatorial confirmatória. A escala demonstrou ajuste adequado do modelo de medição aos dados (erro quadrático médio da raiz <0,06; resíduo quadrático médio padronizado da raiz <0,06) após a exclusão de vários itens. Ao comparar o instrumento ER5FP com o instrumento Inventário Reduzido de Cinco Grandes Fatores de Personalidade (IGFP-5R), com 32 Itens, foi encontrada evidência moderada de validade convergente para extroversão, neuroticismo e abertura à experiência (correlações brutas que variam de 0,44 e 0,57 e correlações desatenuadas de 0,60 a 0,80). Para concordância e consciência, foram encontradas

evidências mais fracas (correlações brutas de 0,33 e 0,29 e correlações desatenuadas de 0,48 e 0,43, respectivamente). Isso significa que o instrumento ER5FP apresenta qualidade psicométrica adequada para aplicação nesse trabalho.

6.3.2 A rede social Wedy

Como alternativa as limitações de acesso a conjuntos de dados públicos e rotulados, esta pesquisa é realizada em parceria com o setor privado. A empresa escolhida como parceira no desenvolvimento da pesquisa é a Wedy, que faz parte do Parque Tecnológico da Unisinos (Tecnosinos). Ela é uma rede social com funcionalidades específicas para a organização de casamentos e foi cofundada em 2014 (ainda com o nome de Mecasei.com) por um dos autores da pesquisa (Daniel Tamiosso). Logo no primeiro ano de fundação, foi selecionada como uma das *startups* nacionais mais relevantes de 2015 (VENTIUR, 2015). A aplicação de conceitos de Inteligência Artificial em um mercado tradicional acabou referenciando a empresa como destaque nacional em publicações da mídia e sendo escolhida pela IBM Global como uma das seis *startups* mais inovadoras do mundo em 2016 (STARTUP-BRASIL, 2016), recebendo em 2017 acesso à capital de risco (STARTSE, 2017). Em 2018, a Mecasei.com iniciou uma estratégia de globalização da marca para a América Latina (com o nome de Wedy) e o avanço rápido de pesquisa e desenvolvimento de produto com tópicos avançados de Ciência da Computação (EXAME, 2018).

De forma colaborativa e online, a rede social da Wedy cria a oportunidade de seus usuários (noivas e noivos) encontrarem o local da cerimônia, criar a lista de presentes, contratar serviços de fotografia, decoração, vestido de noiva e tudo o que envolve a jornada de descoberta, planejamento, organização e divulgação de um casamento. Em termos de conteúdos sociais, os usuários publicam experiências, dúvidas, emoções, produtos e inspirações, e recebem *feedbacks* sociais a partir de curtidas e comentários nas suas publicações. O produto, como pode ser visto na Figura 14, possui como recurso central uma linha do tempo. Atualmente (2021), a Wedy possui 2.860 usuários ativos mensais nessa linha do tempo. No momento de escolha da Wedy como objetivo de estudo, antes das restrições devido à pandemia do coronavirus, detalhadas na Seção 6.4.1, a rede social possuía a tendência de ter uma amostra de dados coletados de mais de 18.000 usuários com seus traços de personalidade inferidos.

A linha do tempo, recurso social da Wedy, utiliza-se de algoritmos de aprendizado de máquina para recomendação dos seus conteúdos. Esse conteúdo – personalizado para cada usuário – está estruturado em cartões, onde cada cartão consiste em uma publicação de um usuário, com questões relacionadas ao domínio do casamento, como inspirações, recomendações de fornecedores, produtos para venda (exemplo na Figura 15), dúvidas específicas, apoio emocional e outros conteúdos relevantes a esse nicho, utilizando-se de texto e imagem como mídia. Cada cartão é um organismo de rede social e uma oportunidade para os usuários socializarem. Por exemplo, um cartão tem como potenciais oportunidades de interações: 1) o usuário indicar que

gostou do seu conteúdo, 2) o usuário responder o conteúdo do cartão com uma mensagem específica, 3) o usuário compartilhar esse conteúdo em outras redes sociais populares, 4) o usuário ver e expandir o conteúdo do cartão, 5) o usuário comprar o produto inserido no cartão, 6) o usuário ver o conteúdo relacionado com esse cartão e navegar para outros cartões e 7) o usuário ver e navegar nas *tags* geradas a partir de Processamento de Linguagem Natural e Visão Computacional e descobrir novos conteúdos.

A partir desse contexto de rede social, a Wedy coletou os dados necessários (personalidade, comportamento e demografia) com o consentimento de seus usuários para essa pequisa e disponibilizou-os de forma anonimizada. Esses dados foram capturados a partir dos materiais descritos na Seção 6.4. De qualquer forma, existem restrições nessa rede social: a) os usuários têm um período de tempo limite dentro da rede, eles estão ativos de 6 a 18 meses (esse é o tempo médio de organização de um casamento pela plataforma); b) não há fatores de conexão entre os usuários (amizade, interesse direto e outras formas de conexões de nós). A falta de conexão direta entre seus usuários ocorre por limitações sociais do contexto de eventos de casamento, onde as amizades de pessoas próximas passando pelo mesmo momento na vida são raras. Essas características específicas dessa rede social devem ser consideradas quando comparar essa análise com análises realizadas para outras redes sociais online.

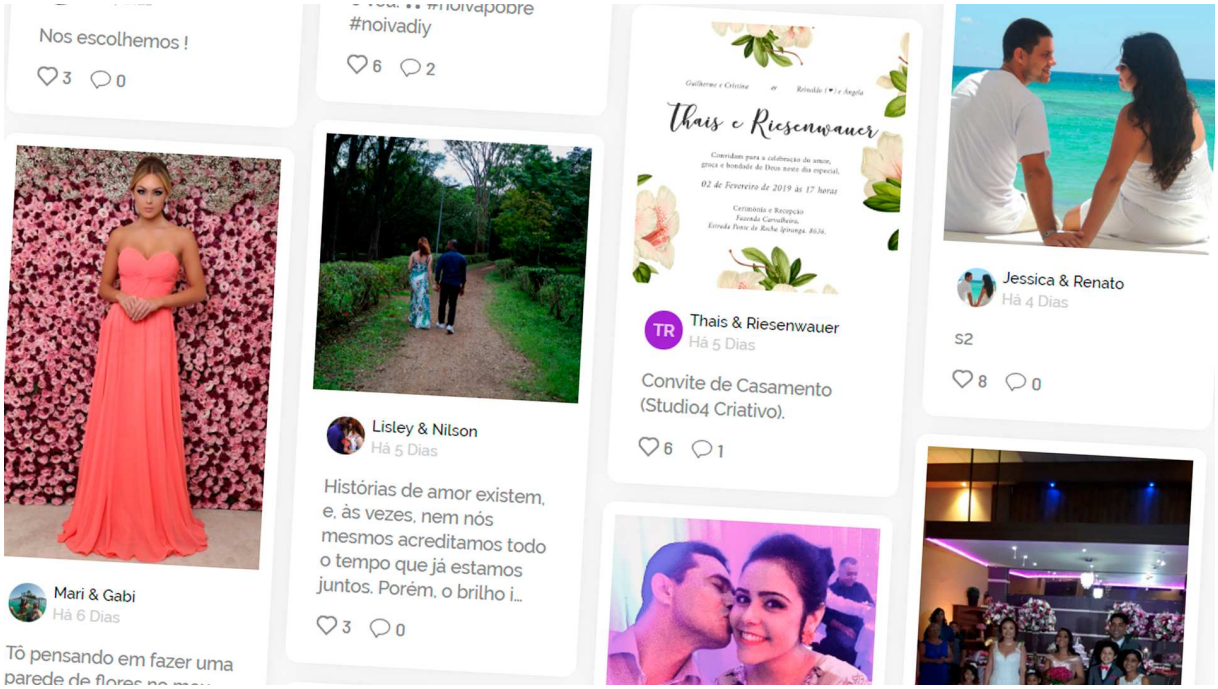

Figura 14: Wedy e sua linha do tempo social

6.4 Conjunto de Dados

O desafio da pesquisa em larga escala da personalidade em redes sociais e, especificamente, do agrupamento de personalidade em conjunto com pegadas digitais com alta dimensionali-

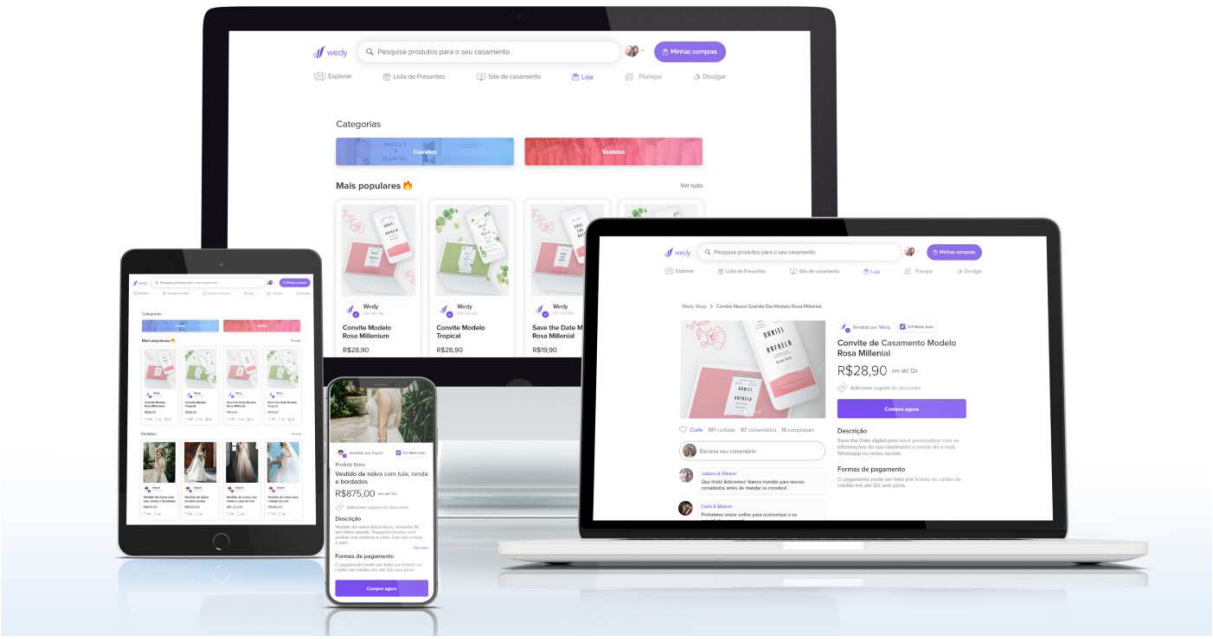

Figura 15: Wedy Marketplace - uma rede de transações financeiras com interações sociais

dade está restritamente relacionado à falta de conjuntos de dados rotulados e à dificuldade de construção de novos conjuntos, seja devido às restrições de privacidades estudadas ao longa da revisão dos trabalhos relacionados ou do alto custo de construção deles. Os poucos conjuntos de dados existentes sofrem com outras deficiências relacionadas ao não anonimato (o que torna os usuários mais relutantes em expressar sua verdadeira personalidade) ou mesmo com expressividade limitada (por exemplo, no Twitter), resultando em conjuntos observacionais pequenos, como foi observado na Tabela 1.

O conjunto de dados revisado a seguir - separado em três conjuntos principais (personalidade, comportamento e demografia), possui uma alta dimensionalidade. Isso acontece porque os dados disponibilizados pela plataforma não incluem somente as pegadas digitais deixadas pelos usuários de forma ativa, como as interações públicas em publicações na linha do tempo. O conjunto de dados também inclui diversos atributos referentes às pegadas digitais passivas, como a frequência de uso, as publicações visitadas e consequentemente os tópicos de interesse relatados apenas pela navegação de cada usuário na rede social. Por outro lado, deve-se atentar às outras características desse conjunto de dados, que de alguma forma podem criar ruídos na análise descritiva de dados. Por exemplo, é percebido no conjunto de dados coletado para esse trabalho a baixa diversidade de tópicos e um forte viés em relação a tópicos relacionados a eventos de casamentos. Claramente, as características inerentes à pesquisa da personalidade expõem a falta de conjuntos de dados de referência. Isso dificulta o desenvolvimento e o entendimento profundo da relação entre a personalidade e o comportamento de usuários nas redes sociais.

A rede social da Wedy, como foi detalhada na Seção 6.3.2, é estruturada a partir de uma linha do tempo, uma funcionalidade comum das redes sociais. A linha do tempo apresenta uma

série de conteúdos gerados pelos usuários da rede social. Cada usuário possui um perfil, onde seus dados são disponibilizados publicamente ou armazenados de forma privada. Por exemplo, nas redes sociais tradicionais como Facebook, Twitter ou Linkedin, cada usuário pode de forma granular escolher quais dados serão exibidos aos demais membros da rede social. Porém, o fato de não publicar um dado ao público em geral não limita o acesso à ele pelos desenvolvedores da rede social em questão, e esse tipo de dado não foi plausível de coleta nem mesmo pelo projeto *myPersonality* (Seção 5.1.1. Embora cada interação na rede social seja documentada e exibida a partir de um perfil único que representa um casal, os dados gerados são realizados sempre a partir de um usuário, que a rede social armazena privativamente. Isso é essencial para evitar ruídos na análise de comportamento e traços de personalidade. Um casal é formado por duas pessoas, com diferentes pontuações nos cinco traços de personalidade.

Conforme pode ser visto na Figura 16, a rede social da Wedy produz cerca de 350 publicações em média, por mês, na sua linha do tempo. Essas publicações são geradas espontaneamente por seus usuários. Além dessas publicações, empresas conectadas à rede social da Wedy publicam aproximadamente 280 produtos por mês na linha do tempo, como mostra a Figura 17. Ambos os tipos de publicações possuem as mesmas opções de micro-interações sociais: curtidas, comentários e visualizações. Mensalmente são cerca de 760 curtidas realizadas pelos usuários da rede (Figura 18) e 150 comentários publicados (Figura 19). De maneira geral, essas são as pegadas digitais ativas visíveis das micro-interações social. Porém, a densidade de dados é maior nas interações que não são publicadas a partir da pegada digital passiva que representa a visualização de um conteúdo. Mensalmente, cerca de 5.200 visitas únicas em publicações disponibilizadas na linha do tempo são realizadas, conforme mostra a Figura 20.

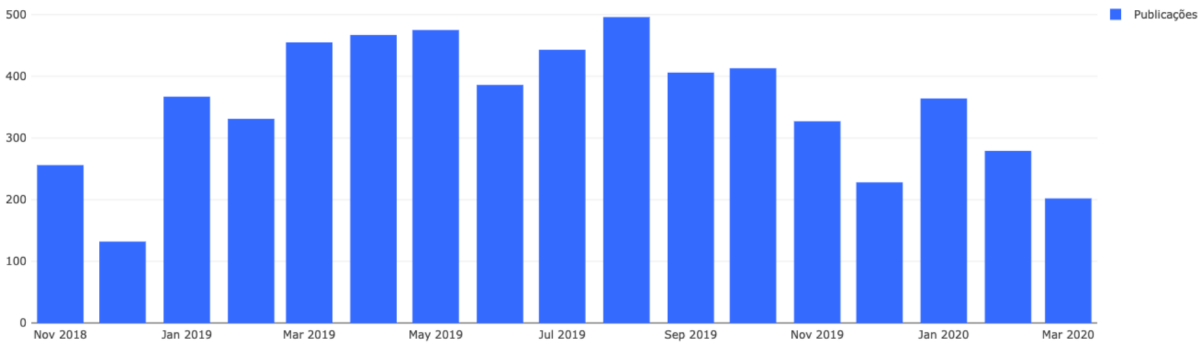

Figura 16: Volume de publicações na linha do tempo da Wedy

6.4.1 Limitações no conjunto de dados

Os dados apresentados nos gráficos da Seção 6.3.2, sobre o volume das principais dimensões de dados da rede social Wedy, estão limitados para o período de de Novembro de 2018 a Março de 2020, início da pandemia da Covid-19 no Brasil. Na época, de acordo com a Organização Mundial da Saúde (OMS), os casos confirmados da Covid-19 já haviam ultrapassado 200 mil

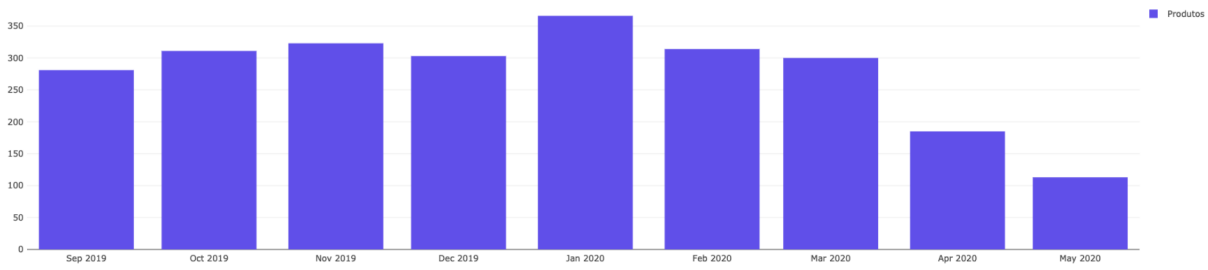

Figura 17: Volume de produtos anunciados na linha do tempo da Wedy

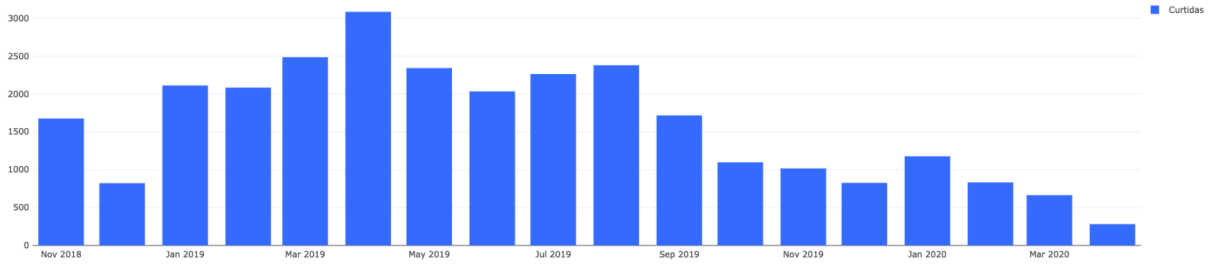

Figura 18: Volume de curtidas nas publicações na linha do tempo da Wedy

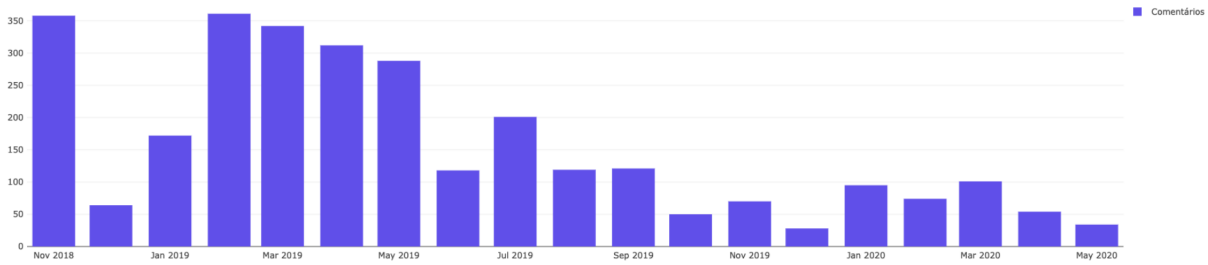

Figura 19: Volume de comentários nas publicações na linha do tempo da Wedy

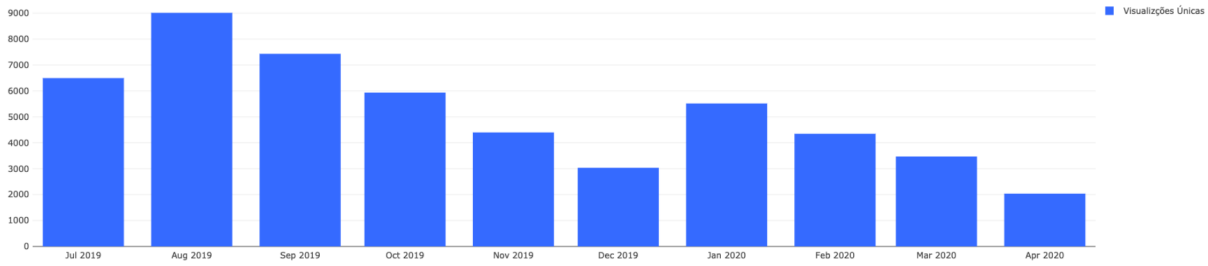

Figura 20: Volume de visualizações únicas nas publicações na linha do tempo da Wedy

em todo o mundo (FREITAS; NAPIMOGA; DONALISIO, 2020). Como a Wedy é uma rede social de pessoas que buscam a realização de eventos de celebração, envolvendo a aglomeração de grandes grupos de pessoas em espaços fechados, a partir desse momento, a empresa sofreu um grande revés no engajamento e na retenção de seus usuários dentro da plataforma. Assim como concertos, reuniões, conferências, esportes e outros encontros familiares, os casamentos foram restringidos das suas realizações desde esse momento até, no mínimo, a conclusão dessa

pesquisa (GÖSSLING; SCOTT; HALL, 2020). Dessa forma, os dados idealizados na proposta da pesquisa, em período prévio à pandemia, a partir do histórico apresentado no Seção 6.4, não foram concretizados. Isso derivou de características do segmento de atuação da rede social, onde os seus usuários utilizam a rede social apenas durante o período de organização do seu casamento. Com as inúmeras incertezas desse período histórico, os casamentos foram colocados em segundo plano, e o volume de dados da rede social caiu +90% em relação ao período anterior, como pode ser observado na Figura 21 .

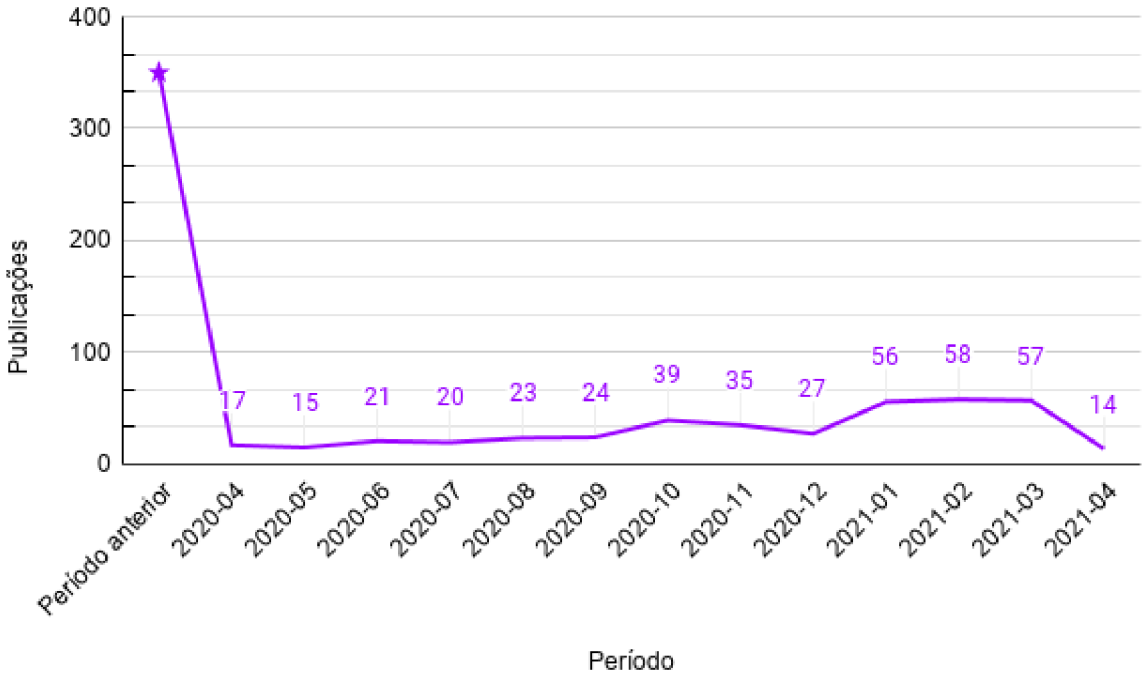

Figura 21: Queda acelerada (+90%) no volume de dados disponibilizados pela rede social Wedy durante a pandemia da Covid-19

Os dados anteriormente à pandemia da Covid-19 disponibilizados pela rede social não foram utilizados na pesquisa, pois aqueles usuários já não estão mais presentes na rede social da Wedy - por atingirem seu objetivo no produto que é a realização dos seus casamentos. Assim, não seria possível obter seu consentimento para a pesquisa, assim como os dados de sua personalidade. Dessa forma, a pesquisa prosseguiu com os novos usuários que entraram no decorrer do processo de coleta de dados da pesquisa, embora em um volume exponencialmente menor do que o previsto na elaboração da pesquisa, e que estão detalhados nas seções posteriores desse trabalho.

6.4.2 Dados de personalidade

Conforme revisado na Seção 2.3, o modelo dos Cinco Grandes Fatores compreende cinco dimensões fundamentais da personalidade humana: abertura à experiência, conscienciosidade,

extroversão, agradabilidade e neuroticismo. Cada uma das cinco dimensões é modelada como valor escalar, cujo valor varia de 4 a 20, representando o grau em que um indivíduo expressa um determinado traço de personalidade ou não. Assim, por exemplo, um valor alto para extroversão indica um indivíduo extrovertido, enquanto um valor baixo para essa dimensão indica uma pessoa introvertida. O conjunto de dados que representa a personalidade dos usuários da rede social estudada foi coletado a partir da distribuição do instrumento psicológico descrito na Seção 6.3.1. O questionário é constituído de 20 questões, onde as respostas seguem uma escala Likert [2] com valores de 1 a 5. As questões perguntavam para o usuário relatar, por exemplo, o quanto ele gosta de cooperar com outras pessoas. Então, o usuário selecionava uma única resposta entre as opções: discordo totalmente (1), discordo (2), nem concordo nem discordo (3), concordo (4) e concordo totalmente (5). Haviam quatro questões para cada um dos cinco traços de personalidade. A soma das respostas dessas quatro questões resulta no índice que representa um traço de personalidade, sendo esse índice variando de 4 à 20. Assim, quanto mais próximo de 20, mais forte é a presença do traço e, por outro lado, quanto mais próximo de 4, mais fraca é a presença do traço. O resultado para cada traço é normalizado, sendo gerada uma escala decimal com valores entre 0,2 e 1. A Figura 22 demonstra como o conjunto de dados de traços de personalidade é representado visualmente aos usuários que respondem ao questionário.

6.4.3 Dados comportamentais

O conjunto de pegadas digitais, ativas ou passivas, deixadas pelos usuários ao utilizarem uma rede social representam uma série de atividades sociais que ocorre de forma online. Dessa forma, membros de uma rede social frequentemente desempenham papéis distintos que podem ser deduzidos a partir de observações das suas atividades online. Especificamente na rede social Wedy, as pegadas ativas e passivas (introduzidas anteriormente nessa seção), representam o comportamento social de usuários que estão planejando e organizando um evento sentimental como é o casamento. Na rede social da Wedy, essas atividades estão concentradas em publicações. Uma publicação é composta por uma série de atributos, tais como:

1. **Conteúdo**: conteúdo da publicação realizada pelo usuário, podendo ser composta de texto e/ou imagem;

2. **Tipo da publicação**: uma publicação pode ser do tipo "Inspiração"quando o conteúdo é criado diretamente pelos usuários para compartilhar algum conteúdo social, ou do tipo "Produto"quando o objetivo é uma publicação de venda de um produto;

3. **Autor da publicação**: quem é o criador da publicação, podendo ser um usuário, representando um noivo ou uma noiva, ou uma empresa que está ofertando seus produtos ou

[2]Likert é o tipo de escala de resposta psicométrica usada habitualmente em questionários, e é a escala mais usada em pesquisas de opinião. Ao responderem a um questionário baseado nesta escala, os perguntados especificam seu nível de concordância com uma afirmação.

serviços;

4. **Tags**: conjunto de marcações realizadas automaticamente a partir de Processamento de Linguagem Natural e/ou Visão Computacional e, também, adicionadas manualmente pelos usuários;

5. **Curtidas**: conjunto de ações realizadas para expressar publicamente que um usuário gostou de uma publicação, onde uma curtida representa uma ação de um usuário em uma publicação em determinado momento de tempo;

6. **Comentários**: conjunto de interações sociais nas publicações a partir de mensagens de textos deixadas por outros usuários ou pelo autor da publicação;

7. **Curtidas em comentários**: conjunto de curtidas realizadas pelos usuários da rede social em relação a um comentário em específico;

8. **Visualizações**: conjunto de pegadas digitais passivas, referentes ao ato de navegação de um usuário em conteúdos da rede social, representado pela ação de um usuário detalhar o conteúdo de uma publicação em um determinado momento no tempo;

9. **Data da publicação**: representa o momento no tempo em que a publicação foi publicada na rede social;

10. **Dados do produto**: quando a publicação é referente a um produto ou serviço, dados adicionais são representados, como características do produto e preço.

Um conjunto de dados experimental foi disponibilizado pela Wedy, entre as datas de 01/07/2019 e 01/03/2019, contendo um *corpus* de 2846 publicações – apenas referentes a publicações do tipo "Inspiração". A Figura 24 representa uma publicação desse tipo realizada dentro da rede social. Nesse conjunto de dados, após a realização de uma análise descritiva, observa-se na Figura 25 e nas demais análises quantitativas, as seguintes estatísticas descritivas que resumem a tendência central, dispersão e forma da distribuição desse conjunto de dados:

- Cada publicação recebe em média 12,73 visitas únicas, onde a publicação mais popular da rede social recebeu 1.383 visitas únicas. Entretanto, a mediana de visitas únicas a uma publicação é de apenas 6;

- Uma publicação em média é composta em 87% dos casos por publicações com imagens;

- A publicação mais popular obteve 69 comentários, porém as chances de uma publicação receber um comentário na rede social é próxima a zero;

- O número de curtidas, que exige um investimento menor do usuário, representa 345 como número máximo de ações desse tipo em uma publicação da rede. Sendo a média de curtidas em uma publicação de 3,60 e a mediana de 2;

- As publicações foram geradas por 1.572 usuários distintos, o que representa uma média de 1,81 publicações por usuário;

- O conjunto de dados, em sua forma bruta, apresenta 49.748 *data points* apenas em relação às dimensões de visualizações, curtidas, imagens e comentários.

No que se refere aos dados comportamentais, especificamente, uma série de novas dimensões observadas em trabalhos relacionados podem ser produzidas a partir do conjunto bruto de dados, tais como:

1. **Frequência de acesso à rede social**: fator que mensura a frequência mensal de acesso a rede social;

2. **Análise linguística de todo conteúdo publicado**: potencial aplicação de diversos algoritmos de análise linguística, tais como LIWC, MRC, *Speech Tags*, *N-Grams*, *Affective Parameters* e outros, bem como dados como o uso de *emoticons*, tipos de frases, uso ou não de letras maiúsculas e tópicos extraídos de marcações (ou *tags*) do conteúdo (texto e imagem).

6.4.4 Dados demográficos

A demografia é definida como dados estatísticos sobre as características de uma população, como idade, sexo e renda das pessoas na população. Por exemplo, quando o censo reúne dados sobre a idade e o sexo das pessoas, este é um exemplo de reunião de informações demográficas. Uma população representada por uma rede social também pode ter sua demografia explorada e coletada no formato de atributos. Esse é o caso do estudo de Filippova (2012), que de forma inversa, desenvolveu algoritmos para predição de atributos de demografia em uma rede social (Youtube) com precisão acima de 90%.

Dessa forma, como um conjunto de dados com potencial de relevância para os objetivos desa pesquisa, os principais atributos demográficos da rede social da Wedy são os seguintes:

1. **Data de inscrição**: representa o momento (data e horário) que o usuário realizou o cadastro na rede social;

2. **Data do casamento**: atributo que representa o momento mais relevante para os usuários dentro da rede social;

3. **Tamanho do casamento**: um casamento dentro da rede social é representado pelo seu tamanho, onde os dados disponibilizados pela rede social estão previamente discretizados de acordo com o número de convidados, onde pequeno é até 50 convidados, médio entre 50 e 200 convidados e grande com mais de 200 convidados;

4. **Investimento no casamento**: para cada cidade do país um índice é calculado pela rede social de acordo com o valor médio estimado de um casamento. Quando o valor se encontra abaixo desse intervalo médio, o investimento é baixo, quando encontra-se exatamente no intervalo médio é considerado um casamento de investimento médio e aqueles casamentos que extrapolam o intervalo médio são considerados casamentos de alto investimento;

5. **Estilo do casamento**: um casamento pode ser realizado seguindo diversos estilos, representando os desejos e a personalidade de cada usuário, por exemplo clássico ou tradicional, rústico, retrô ou vintage, temático ou personalizado, moderno ou descolado, sóbrio e outros;

6. **Tempo de planejamento do casamento**: a diferença em dias entre a data de casamento e a data de cadastro de um usuário.

6.5 Método

Esta seção descreve o método utilizado para o desenvolvimento do trabalho. Observando o conjunto dados descritos na Seção 6.4, estimou-se o potencial de aproximadamente 1.500 participantes, devido ao mesmo período de tempo entre o experimento a ser realizado e o conjunto de dados experimental analisado anteriormente. Porém, como revisto anteriormente, a coleta dos dados na rede social alvo do estudo sofreu um forte impacto devido ao seu contexto de atuação (eventos), a sua relação com pandemia do COVID-19 e as políticas governamentais de distanciamento social, como descrito na Seção 6.4.1. Dessa forma, a pesquisa está restrita a 1/10 do potencial projetado de participantes, onde apenas dados de 157 participantes foram selecionados para o estudo. As próximas seções introduzem e descrevem o passo-a-passo do trabalho com esse conjunto de dados.

6.5.1 Coleta de dados

Pode-se observar na Tabela 3 uma visão geral dos dados disponibilizados pela Wedy. São três matrizes de dados relativamente esparsas contendo aproximadamente 450 colunas, com a combinação dos dados descritos na Seção 6.4. A primeira matriz contem dados de pegadas digitais ativas e passivas que representam os dados comportamentais. A segunda matriz contém os dados demográficos. A terceira matriz representa os dados extraídos a partir da aplicação do instrumento psicológico de inferência de traços de personalidade descrito na Seção 6.3.1.

6.5.2 Pré-processamento dos dados

O conjunto de dados oferecidos para a pesquisa pode conter uma série de inconsistências e redundâncias, sendo na sua essência imperfeitos para o desenvolvimento imediato desse tra-

Tabela 3: Uma visão geral das características dos conjuntos de dados

Conjunto de dados	Características
Comportamental	(1) Publicações: (a) Conteúdo, (b) Tipo da publicação, (c) Categorias de conteúdos visitados, (d) Comentários, (e) Total de comentários recebidas, (f) Total de likes recebidos, (g) Data; (2) Curtidas: (a) Publicação; (3) Comentários: (a) Conteúdo, (b) Publicação; (4) Visitas: (a) Conteúdo (5) Total de Conteúdos publicados (6) Total de Conteúdos consumidos
Demográfico	(1) Data de inscriçao (2) Data de casamento (3) Frequência de uso mensal (4) Tamanho do casamento (5) Investimento no casamento (6) Estilo do casamento (7) Tempo de planejamento do casamento
Personalidade	(1) Cinco Grande Fatores de Personalidade: (a) Abertura para a experiência, (b) Conscienciosidade, (c) Extroversão, (d) Neuroticismo ou Instabilidade Emocional, (e) Amabilidade

balho. Para prosseguir com o desenvolvimento dessa pesquisa, técnicas de pré-processamento descritos por García et al. (2016) são utilizadas para remover os dados ruidosos ou para imputar (preencher) os dados ausentes:

1. **Limpeza dos dados**: nesse processo é analisado o conjunto de dados a procura de informações redundantes ou irrelevantes para os objetivos dessa pesquisa.

2. **Imputação de valores ausentes**: esse processo evita a suposição que o conjunto de dados está completo, pois nessa etapa é analisado valores incompletos ou ausentes que potencialmente podem ser substituídos por valores informados manualmente;

3. **Discretização**: O tipo dos dados presente no conjunto pode variar, desde categóricos, onde nenhuma ordem entre os valores podem ser estabelecidos, para dados numéricos onde a ordem entre os valores existentes. Dessa forma, a tarefa de discretização é útil nesse estudo para melhorar o desempenho e eficácia de clusterização ao reduzir conjuntos de dados de valor contínuo.

6.5.3 Clusterização aplicada

A partir da conclusão de coleta de dados e do pré-processamento dos dados, o primeiro objetivo do trabalho nessa fase é encontrar conjuntos de atividades que comumente ocorram com mais frequência e agrupá-las em categorias que representem os comportamentos dos usuários na rede social em combinação com os dados de personalidade e para os dados demográficos previamente normalizados, reduzidos e selecionados como características relevantes. Para isso, o trabalho utiliza-se de estratégias de agrupamento de aprendizado de máquina não supervisionado, ao invés de tentar corresponder comportamentos a um conjunto predefinido por observação humana.

Uma abordagem comum para extração de dados para a criação de agrupamentos dinâmicos (sem conhecimento prévio das características potenciais de agrupamento), também revisada na Seção 4, é a aplicação do algoritmo *K-means*. Embora essa seja a técnica inicial sugerida para a criação dos agrupamentos, outras técnicas serão discutidas para o desenvolvimento de modelos comparativos de análise, como Spectral Clustering, K-medoids e Agglomerative Clustering. Dessa forma, este trabalho identifica como a combinação dos três conjuntos de dados descritos

na Seção 6.4 agrupam-se em perfis de usuários ativos de redes sociais para posterior análise descritiva e validação do método, descritos na próxima seção.

6.5.3.1 Validação da clusterização

Uma das principais considerações em relação ao agrupamento dos dados é selecionar o número certo (indicado por k) de *clusters* a serem extraídos. Infelizmente, não existe uma maneira correta (ou simples) de fazer isso. Além disso, o valor desejável de k depende da aplicação pretendida. Se o objetivo é obter informações a partir dos dados, um pequeno número *clusters* pode ser mais fácil de interpretar e visualizar. Por outro lado, se o objetivo é construir modelos preditivos, um número maior de agrupamentos reterá mais informações da matriz original, permitindo previsões mais precisas. Por isso, uma vez formados os agrupamentos, os resultados foram avaliados através de dois critérios de avaliação já revisados na Seção 4, ou seja, *Davies–Bouldin Index* (DBI) e *Silhouette Coefficient* (SC). O DBI calcula a relação entre *cluster* e *intra-cluster* e o SC verifica a semelhança de cada objeto com todos os outros objetos em seu próprio *cluster* e sua dissimilaridade com objetos pertencentes a outros agrupamentos.

6.5.4 Análise descritiva

Os agrupamentos extraídos dos conjuntos de dados subjacentes potencialmente resultam em grupos coerentes, onde indivíduos do mesmo grupo têm características semelhantes. Os dois índices de validade de *cluster*, ou seja, DBI e SC, fornecerão a medida de quão bem separados esses agrupamentos estarão um do outro, além da coesão interna do agrupamento. No entanto, para estudar o que representa as propriedades de cada agrupamento, é necessária uma análise descritiva. Geralmente, a análise descritiva é baseada no conjunto de recursos utilizados no agrupamento, ou seja, todos os três conjuntos de dados e as características (brutas e/ou produzidas) selecionadas de cada um desses conjuntos. É nessa etapa da pesquisa que estudou-se a viabilidade da formação de grupos coerentes, verificando-se a coesão dos indivíduos de um mesmo grupo a partir do intervalo de semelhança de suas características analisadas por diferentes algoritmos e distintas parametrizações, bem como seleção de características distintas. O objetivo é observar quais combinações comportamentais, demográficas e socioafetivas são distribuídas entre os grupos criados e a importância de cada dimensão na formação dos agrupamento, observando os seus dados descritivos como intervalos de valores, médias, bem como generalizações contextuais relacionadas ao estudo da personalidade como a combinação de determinados traços de personalidade com perfis comportamentais e demográficos dos usuários segmentados em cada agrupamento.

6.5.5 Questões de privacidade

O esforço essencial de reunir e usar traços de personalidade eticamente seguiu as diretrizes gerais de outras pesquisas científicas comportamentais de consumidores, funcionários ou pacientes. Eles incluem: transparência de intenção e uso; cumprimento das leis e regulamentos de privacidade; e alinhamento dos interesses do pesquisador com os dos usuários (JACHIMOWICZ; MATZ; POLONSKI, 2017).

Esse último princípio foi o ponto de partida dessa pesquisa. Embora o estudo possa apontar características relevantes do ponto de vista comercial para a empresa parceira nesse estudo, as informações futuramente publicamente por essa pesquisa, a partir de dados anonimizados, estão diretamente relacionados ao objetivo de entender o comportamento de usuários de redes sociais e futuramente potencializar os benefícios e a experiência de uso à eles. Dessa forma, essa pesquisa comprometeu-se a tornar transparente o objetivo dessa pesquisa e a relevância para a coleta das pegadas digitais, o cumprimento das legislações e regulamentos de política de privacidade e criar um alinhamento conjunto entre pesquisadores, empresa parceira e usuários da rede social no desenvolvimento da pesquisa. Além desses compromissos, essa pesquisa não tem como objetivo alterar ou experimentar novas funcionalidades na rede social estudada, mas sim, apenas entender como se comportam usuários de perfis diferentes dentro da rede social sem interferir na sua experiência e muito menos expor seus dados publicamente.

Figura 22: Exemplo de resultado apresentado ao usuário após envio das respostas do questionário. Esses dados poderão ser adaptados ao contexto inerente à rede social.

Figura 23: Exemplo de etapa de coleta de dados de personalidade a partir de formulário online.

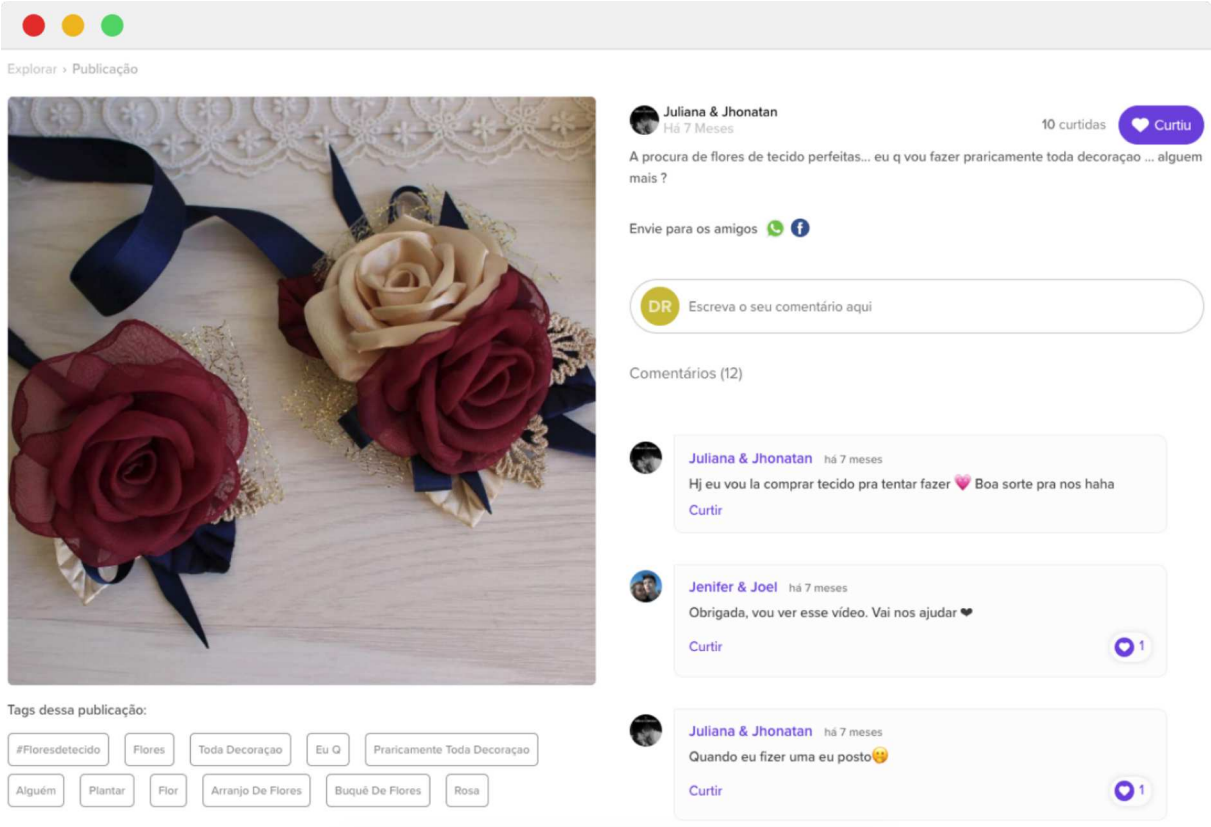

Figura 24: Exemplo de uma publicação na rede social da Wedy, quando acessada por um navegador de Internet de um computador.

	Visualizações	Curtidas	Publicações com Imagens	Comentários
Média	12.73	3.60	0.87	0.28
Desvio Padrão	33.21	7.93	0.37	1.67
Mínimo	1.00	0.00	0.00	0.00
25%	3.00	1.00	1.00	0.00
50%	6.00	2.00	1.00	0.00
75%	13.00	4.00	1.00	0.00
Máximo	1383.00	345.00	3.00	69.00

Figura 25: Estatísticas descritivas do conjunto de dados comportamentais dos usuários da rede social Wedy.

7 ANÁLISE E RESULTADOS

Esta seção descreve, o processo de mineração de dados, os experimentos realizados e o resultados obtidos. A coleta de dados, o pré-processamento de dados, bem como a seleção de características são explicadas primeiramente, seguidos pelos experimentos de agrupamento, classificação e análise dos clusters. Os resultados dessas etapas buscam atingir o primeiro objetivo do trabalho, que é o desenvolvimento de agrupamentos considerando pegadas digitais, inclusive pegadas digitais passivas, comportamentais e demográficas, bem como dados socioafetivos (traços de personalidade). Para o segundo objetivo desse trabalho, que refere-se a análise qualitativa e quantitativa dos grupos produzidos, a fim de entender a qualidade dos grupos formados e a validade deles, como o estudo da relação e da relevância de traços de personalidade na formação dos agrupamentos quando colocados ao lado das demais dimensões comportamentais e demográficas, são apresentados os resultados das análises estatísticas entre as variáveis coletadas e processadas.

7.1 Processo de Clusterização

Como essa pesquisa visa a descoberta de padrões a partir de dados comportamentais (pegadas digitais), demográficos e traços de personalidade previamente coletados, foram utilizados algoritmos de aprendizado não supervisionado, e especificamente algoritmos de clusterização. Porém, antes da aplicação de tais algoritmos, um processo inicial de Mineração de Dados foi realizado para formações mais coesas de grupos de usuários de acordo com suas características mais relevantes. Dessa forma, essa seção descreve as etapas iniciais de coleta de dados, pré-processamento de dados e seleção inicial de características.

7.2 Coleta de dados

O primeiro passo para o desenvolvimento do trabalho foi a coleta de dados. A rede social Wedy realiza a coleta e armazenamento de todos os dados comportamentais (ativos e passivos) e demográficos apresentados na Seção 6. Esses dados, disponibilizados em dois conjuntos de dados separados, foram combinados para a obtenção de um único conjunto de dados. Esse novo conjunto de dados, formado pela intersecção do identificador de usuário (presente em ambos os conjuntos de dados iniciais), apresenta pegadas digitais de 16.891 usuários únicos.

Para completar esse conjunto de dados, uma intervenção técnica foi realizada no produto da Wedy. O instrumento de inferência de traços de personalidade, denominado Escala Reduzida de Cinco Grandes Fatores de Personalidade (ER5FP) e descrito na Seção 6.3.1, foi implementado em forma de um questionário amigável de autorrelato de curta duração, seguindo estritamente a metodologia proposta, e disponibilizado online para os usuários ativos[1] na rede social, ao

[1] Usuários que ainda não casaram, e estão realizando pelo menos o seu segundo acesso na rede social

Tabela 4: Análise descritiva do conjunto de dados comportamentais

	total_visits	total_publications	total_likes	total_comments	visits_received	likes_received	comments_received	visits_received	guests_vists	tasks_done	gifts_received
Total	18183.000000	18183.000000	18183.000000	18183.000000	18183.00000	18183.000000	18183.000000	18183.00000	18183.000000	18183.00000	18183.000000
Média	4.602706	0.596546	1.325634	0.132926	4.14959	1.066931	0.109828	4.14959	119.141066	0.07903	2.098664
Desvio Padrão	13.450737	43.220557	30.567245	2.690459	174.30434	10.480554	1.607506	174.30434	536.436382	1.27699	7.135808
Valor mínimo	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000
25%	1.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.00000	0.000000	0.00000	0.000000
50%	2.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.00000	2.000000	0.00000	0.000000
75%	4.000000	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.00000	54.000000	0.00000	0.000000
Valor máximo	635.000000	5825.000000	3341.000000	220.000000	23210.00000	945.000000	118.000000	23210.00000	39987.000000	70.00000	179.000000

Tabela 5: Análise descritiva do conjunto de dados demográficos

	event_budget	event_style	event_type	event_size	event_planning_duration_months
Total	18183.000000	18183.000000	18183.000000	18183.000000	18183.000000
Média	-0.321509	2.304020	1.059341	-0.260628	12.891774
Desvio padrão	1.173828	3.834847	1.999078	1.248558	203.089002
Valor mínimo	-1.000000	-1.000000	-1.000000	-1.000000	-401.833300
25%	-1.000000	-1.000000	-1.000000	-1.000000	3.900000
50%	-1.000000	-1.000000	3.000000	-1.000000	8.566700
75%	1.000000	6.000000	3.000000	1.000000	12.000000
Valor máximo	3.000000	9.000000	3.000000	3.000000	24081.933300

efetuarem o acesso ao produto no seguinte endereço https://app.wedy.com/quiz.

Porém, devido às limitações relatadas ao longo do trabalho e especificamente na Seção 6.4.1, apenas 285 usuários tiveram seus traços de personalizados obtidos pelo formulário online desenvolvido por essa pesquisa. As Tabelas 4, 5 e 6 apresentam uma rápida análise descritiva sobre o conjunto de dados disponibilizados para a pesquisa. A descrição completa do significado de cada uma das dimensões descritas aqui e no decorrer do trabalho podem ser encontradas no Anexo C.1.

Ao analisar de forma isolada os três conjuntos de dados e suas dimensões descritivas, podemos observar a correlação interna de cada um deles pelo coeficiente de Pearson. Na Figura 28, que representa o conjunto de dados comportamental, destacam-se: (a) uma correlação maior, embora ainda fraca, entre o total de conteúdos visitados e o número de curtidas e os comentários deixados e principalmente a quantidade de comentários e curtidas recebidas; (b) a ausência de correlação entre a retenção de usuário (frequência de uso) com a quantidade de conteúdos consumidos; e (c) a correlação moderada entre o número de comentários deixados e o total de curtidas deixadas em conteúdos visitados. Na Figura 27, pode-se observar, analisando o conjunto de dados demográficos, que (a) há uma correlação fraca negativa entre o estilo do evento e o tamanho do casamento; e (b) uma também correlação fraca entre o poder investimento no

Tabela 6: Análise descritiva do conjunto de dados personalidade

	user_id	extroversion	conscientiousness	agreeableness	openness	neuroticism
Total	285	285	285	285	285	285
Média	388616.403509	0.770877	0.844561	0.831404	0.774737	0.789298
Desvio padrão	48156.624286	0.158595	0.115963	0.112923	0.154943	0.164925
Valor mínimo	39757	0.200000	0.200000	0.200000	0.200000	0.200000
25%	386275	0.700000	0.800000	0.750000	0.700000	0.700000
50%	402393	0.800000	0.850000	0.850000	0.800000	0.800000
75%	409784	0.900000	0.900000	0.900000	0.900000	0.900000
Valor máximo	415147	1	1	1	1	1

evento e o tamanho do casamento. Na Figura **??** que representa a análise correlacional das variáveis representando os traços de personalidade, pode-se destacar (a) a correlação moderada positiva com índice mais alto entre abertura à experiências e extroversão; (b) as demais correlações moderadas positivas mais relevantes entre amabilidade e extroversão, conscienciosidade e extroversão; e (c) as correlações ínfimas entre neuroticismo com extroversão, amabilidade e abertura à experiências, ao contrário da relação de neuroticismo com conscienciosidade que apresenta uma correlação positiva moderada.

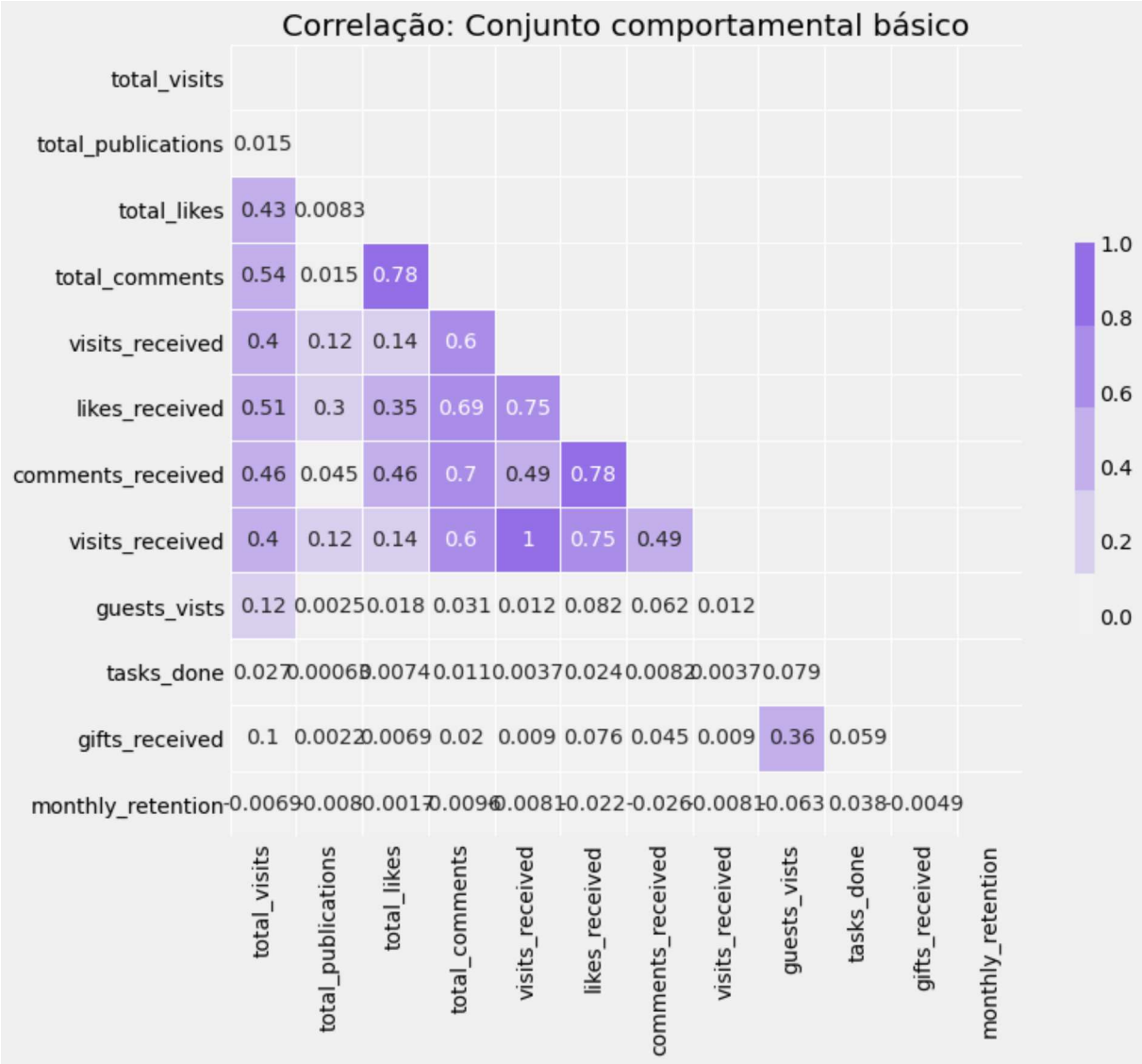

Figura 26: Correlação do conjunto de dados comportamental

O conjunto de dados referentes aos traços de personalidade, coletados manualmente, como observado anteriormente, possui o menor volume de dados para o estudo. Como esses dados são fundamentais para a pesquisa, eles são fator decisivo na restrição do volume total de dados considerados para o estudo, invalidando e removendo do estudo os demais usuários que não tiveram seus dados de personalidade coletados. Ou seja, a interseção dos três conjuntos de dados

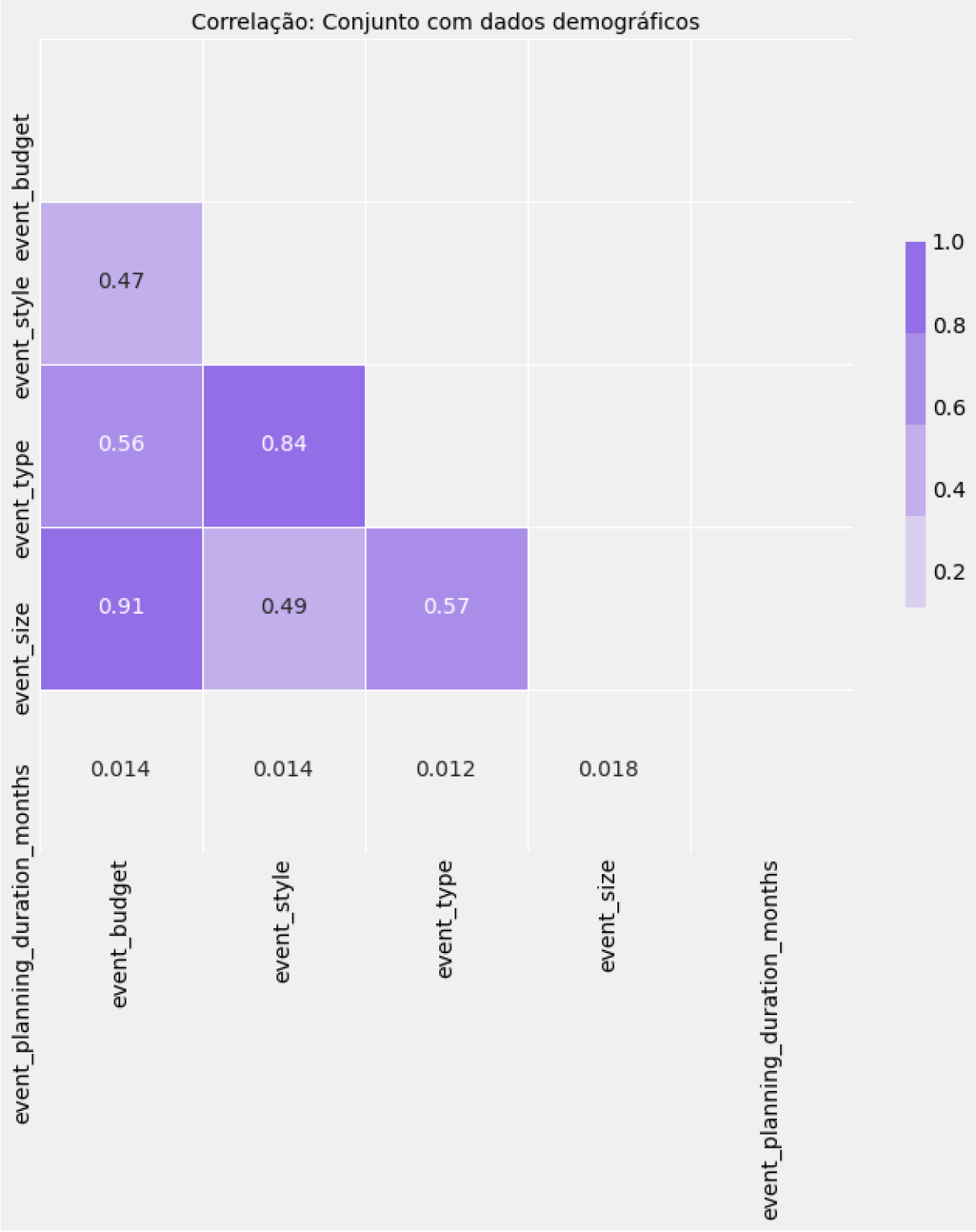

Figura 27: Correlação do conjunto de dados demográficos

está limitada no conjunto de dados de personalidade. Dessa forma, quando combina-se os três conjunto de dados, o volume de usuários estudados está limitado ao número de 285 pessoas. A análise correlacional, apresentada na Figura 29, representa as potenciais correlações entre a interseção desses conjuntos apresentados anteriormente. Entretanto, diretamente, nenhuma

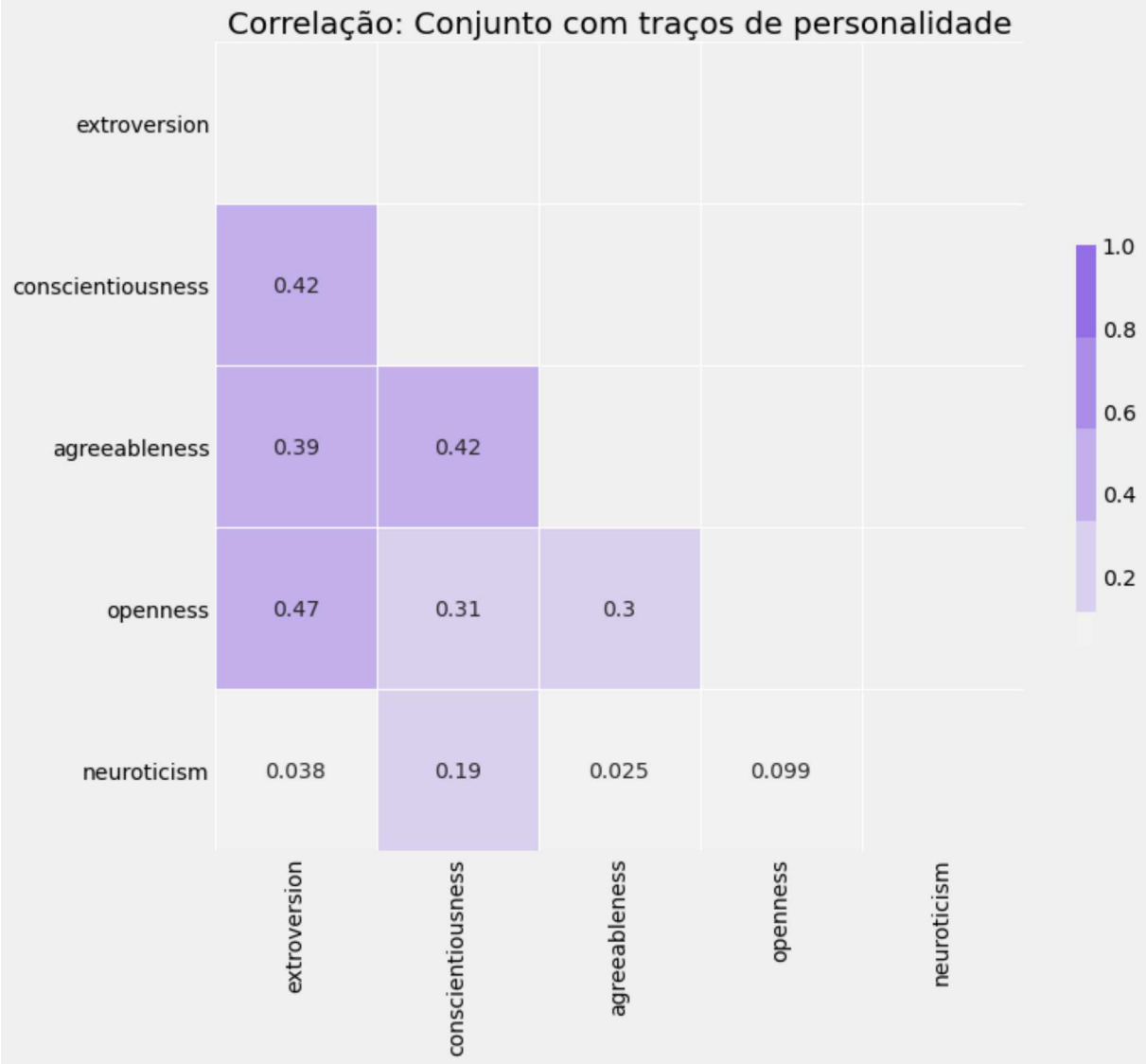

Figura 28: Correlação do conjunto de dados de personalidade

combinação de duas variáveis de conjuntos distintos apresentaram correlações fortes entre elas. Algumas dimensões, como o gênero do usuário, possui correlação fraca negativa entre algumas características de personalidade dos usuários da rede como conscienciosidade, amabilidade e abertura à experiências. A quantidade de curtidas (*likes*) recebidas apresenta uma correlação fraca entre para abertura à experiência e conscienciosidade - a qual também correlaciona-se de forma fraca com o total de publicações realizadas. Na mesma mensuração de correlação positiva e fraca, também encontra-se extroversão e o total de visitas realizadas em publicações de outros usuários, bem como o total de atividades concluídas, correlacionando-se na mesma intensidade com o traço de personalidade de abertura à experiências e características do evento como o seu tamanho e estilo.

Os dados comportamentais são estendidos por mais duas dimensões relevantes aos conteúdos textuais e conteúdos em imagem publicados pelos usuários da rede social. Dos conteúdos especificamente textuais, 3.714 usuários realizaram publicações (Figura 31). Referente aos con-

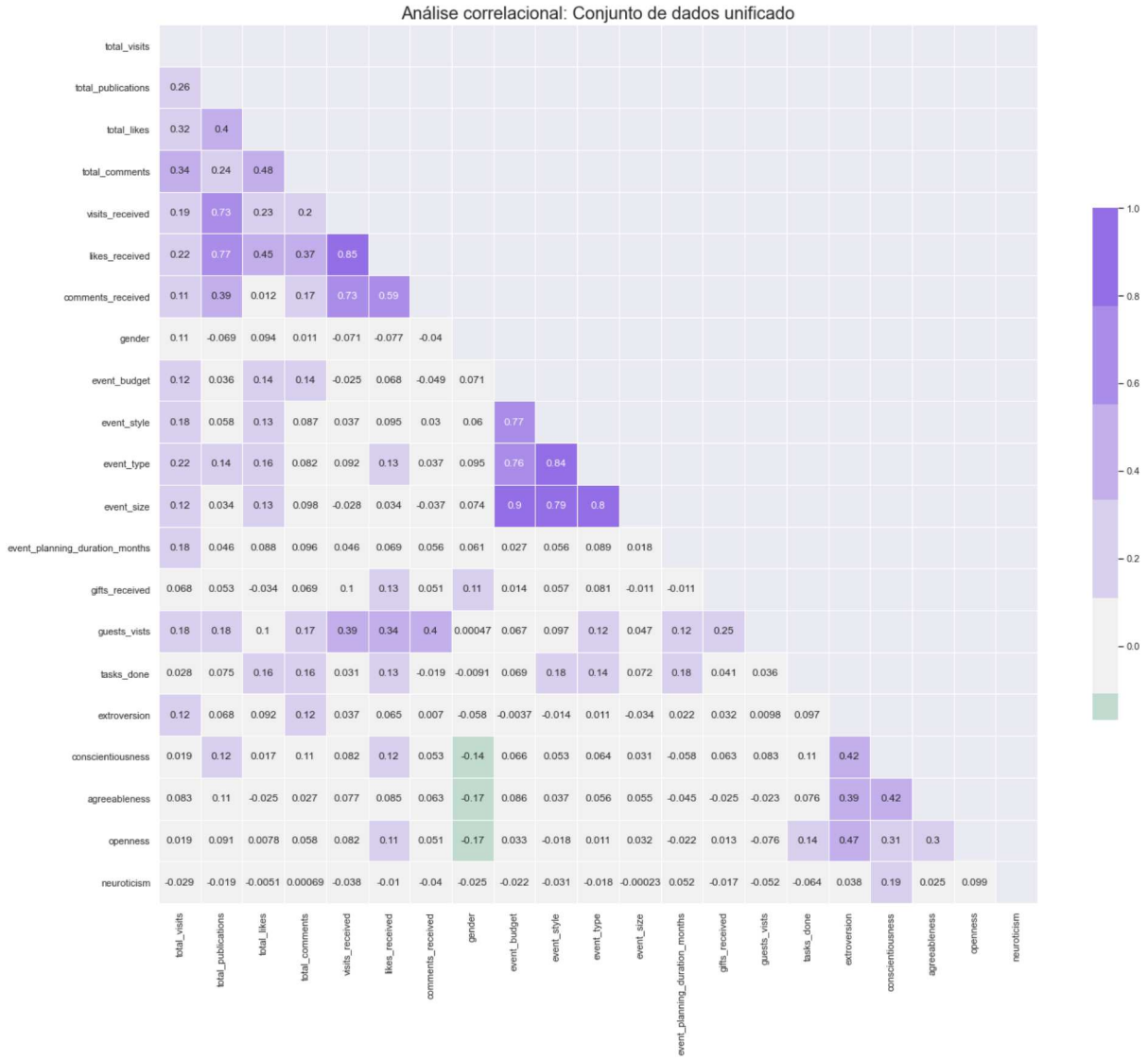

Figura 29: Correlação do conjunto de dados unificado: comportamento, demografia e personalidade

teúdos em mídia (fotos), 3.031 enviaram fotos em suas publicações. Além dessas duas dimensões de dados, outras 416 colunas representam a relação entre um usuário e os seus conteúdos preferidos na rede social Wedy, em uma matriz similar ao exemplificado na Figura 30, com a adição de total de visitas relacionados a cada tópico e não apenas a informação se o usuário demonstrou ou não interesse no tópico, como verificado nos trabalhos relacionados na Seção 5. Na Figura 32 pode-se notar os 50 tópicos frequentemente mais visitados pelos usuários da rede social, coletados a partir de suas pegadas digitais passivas ao navegar entre os conteúdos disponibilizados na rede social, como Vestuário, Convite, Moda, Deus, Amor, Sapato, Bolo e outros. Na Figura 33 estão os tópicos menos frequentes disponibilizados no conjunto de dados como Decote nas Costas, Corpo, Bebê, 1ª Vez, Mãe de Noiva, Pub e outros.

Usuário	FLOR	VESTIDO	SELFIE	NOIVA
AKAM	●	●	●	●
MU	●	●	●	●
FHN	●	●	●	●
LSI	●	●	●	●
ALTR	●	●	●	●

Figura 30: Uma hipotética relação entre um usuário e os seus conteúdos preferidos na rede social Wedy.

Figura 31: Principais palavras publicadas pelos usuários

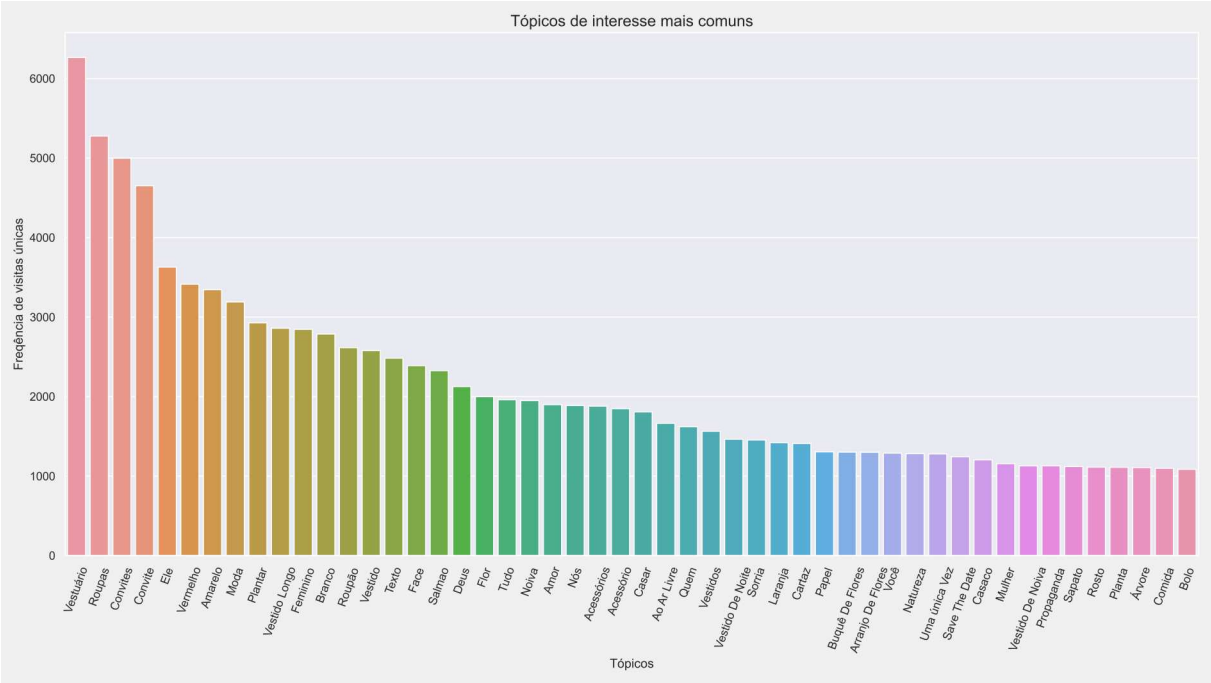

Figura 32: Tópicos de maior interesse coletados a partir de pegadas digitais passivas

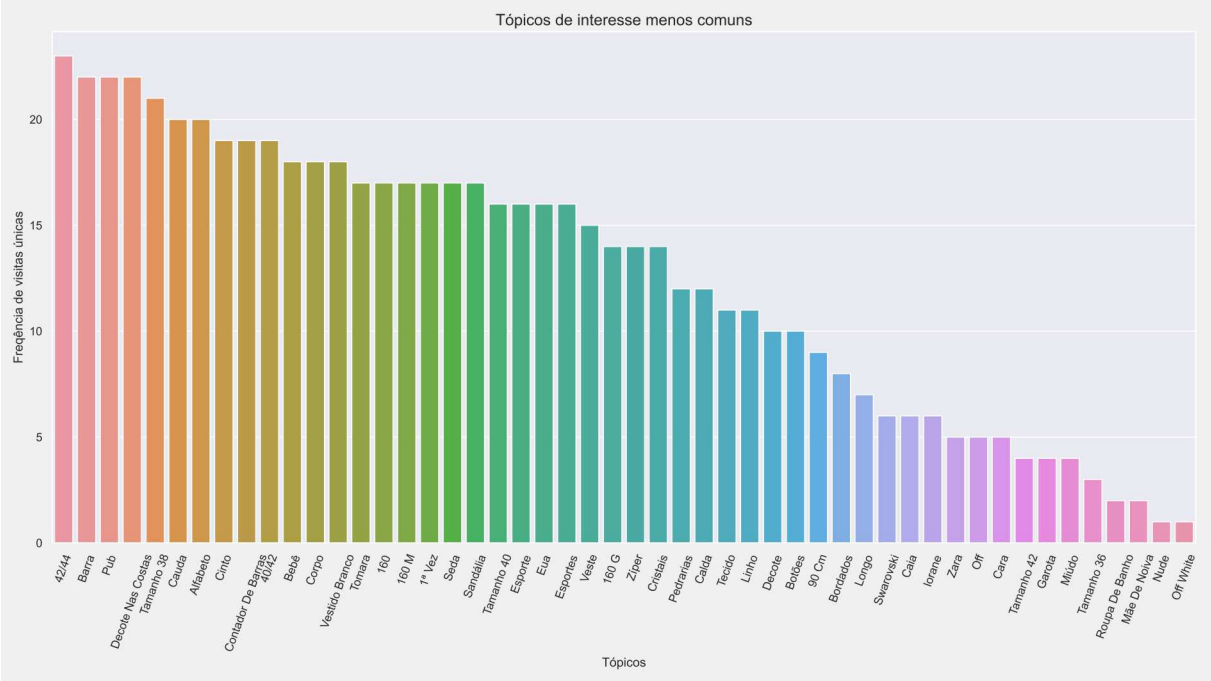

Figura 33: Tópicos de menor interesse coletados a partir de pegadas digitais passivas

7.3 Pré-processamento de dados

Após a fase inicial de coleta, temos informações dos usuários nas três categorias desejadas (comportamental, demográfica e de personalidade), representadas como conjuntos de vetores individuais. No entanto, a maior parte dos dados disponibilizados não possuem dados comportamentais na rede social ou não possuem dados de personalidade. Esses dados não agregam valor aos objetivos dessa pesquisa e podem ser removidos do conjunto. Combinando os três conjuntos de dados e agrupamento eles pelo identificador do usuário, temos acesso à dimensões para cada usuário (Apêndice C.1), resultando em uma matriz total de 16.891 linhas por 447 colunas. Consequentemente ao restringir essa matriz, pelos 188 usuários com traços de personalidade, chegamos ao conjunto de dados final definido pelas dimensões (188,447).

Após a intersecção dos conjunto de dados e criação do conjunto unificado de dados para a continuação do trabalho, verifica-se algumas **características demográficas** relevantes:

- **Gênero**: 91% dos usuários presentes no conjunto de dados possuem gênero informado, onde 88.3% da audiência analisado representam o gênero feminino e apenas 11.7% representam a audiência feminina da rede social (Figura 34);
- **Localização**: 77% dos usuários são residentes da região sudeste do Brasil (Figura 35), e apenas 1% deles representam a a região norte do Brasil;
- **Características dos eventos**: 82% dos usuários que decidiram o quanto investir no seu casamento realizam investimentos baixos ou médios, em casamentos de tamanhos pequenos ou médio em 94% dos casos, sendo que 52% deles optaram por casamentos clássicos, e a totalidade dos que decidiram sobre o tipo do casamento, preferem o seu casamento na cidade, como pode ser observado na Figura 36.

Figura 34: Distribuição por gênero dos usuários da rede social com traços de personalidade coletados

Dessa forma, pode-se concluir que demograficamente a maior parte dos usuários presentes no conjunto de dados combinado são mulheres, residentes na região sudeste brasileira, com um poder de investimento baixo ou médio e que preferem casamentos na cidade no estilo clássico.

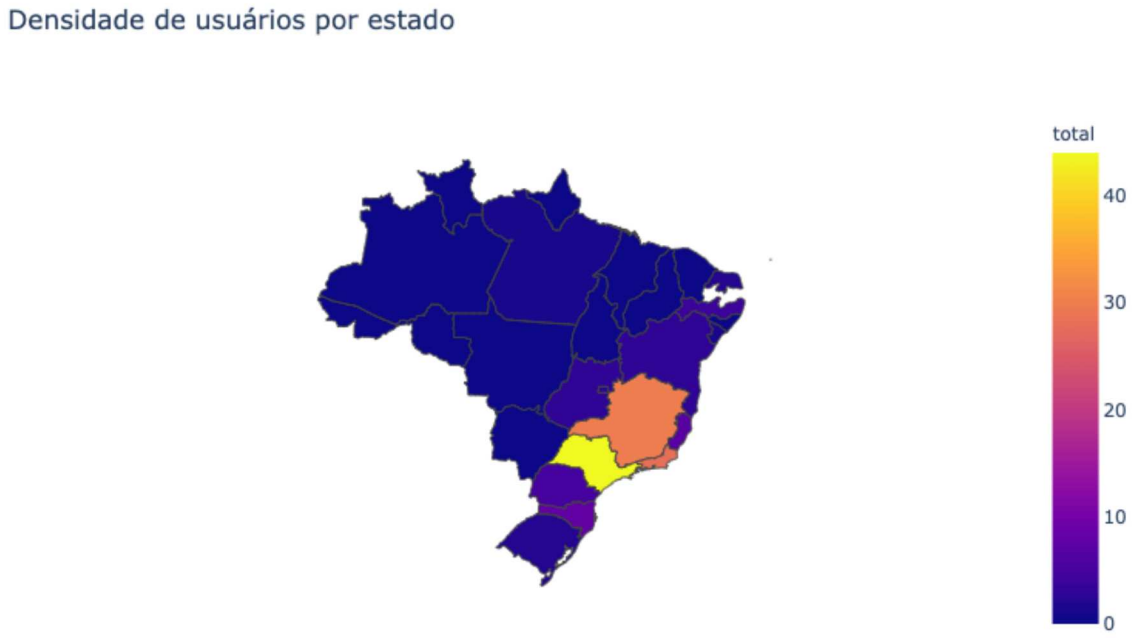

Figura 35: Densidade de usuários por região brasileira com traços de personalidade coletados

Ao analisar especificamente os dados de personalidade desse conjunto, agora restrito exatamente pelos usuários com traços de personalidade inferidos, pode-se observar que os dados estão normalizados percentualmente em relação ao total de pontos máximo que cada traço pode receber via inferência do instrumento ER5FP. Em um comparação simples pelo gênero que o usuário se identifica, pode-se observar uma diferença mais relevante no traço de abertura à experiências dos usuários da rede social, onde a mediana para o gênero masculino é 0.8 e para o gênero feminino 0.65 (Figura 37). Nos demais traços não há diferença relevante na amostragem geral, ao mesmo tempo que há uma maior concentração de valores na metade superior dos vetores de cada traço (Figura 38), principalmente nos traços de conscienciosidade e amabilidade.

Com um conjunto tão restrito de dados, a distribuição de atividades comportamentais através da coleta de pegadas digitais ativas e passivas é notoriamente afetada para além da baixa profundidade de interação devido ao período de análise e suas características macro-econômicas. Na Figura 39, essa distribuição curta se torna visualmente interpretável, onde cada usuário desse conjunto de dados combinado realiza em média 10 visitas únicas em publicações, deixam aproximadamente 2 curtidas nas publicações que interessam e publicam em média 0.44 novos conteúdos.

Após a fase inicial de coleta de dados e a análise inicial das informações dos usuários, representado como um conjunto de vetores reduzido para somente aqueles com traços de personalidade inferidos, pode-se seguir o processo de pré-processamento de dados a fim de remover

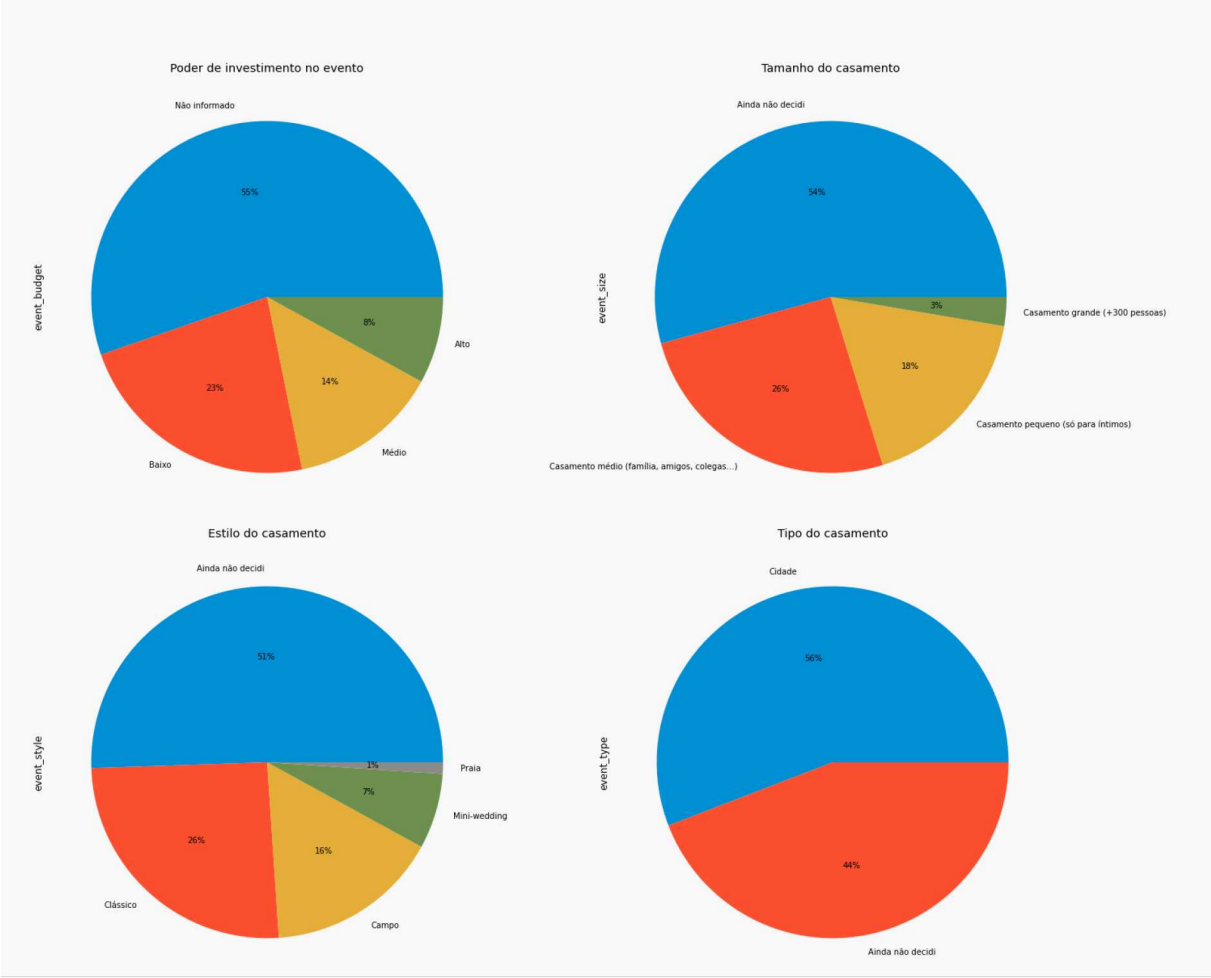

Figura 36: Características dos eventos organizados pelos usuários com traços de personalidade coletados

dados irrelevantes, escalonar dados para evitar que o algoritmo proposto não fique enviesado para as variáveis com maior ordem de grandeza, verificar dimensões com ausência de valores para determinados usuários e considerar técnicas de imputação de valores em dados ausentes e também aplicar a discretização de atributos onde mostrar-se relevante.

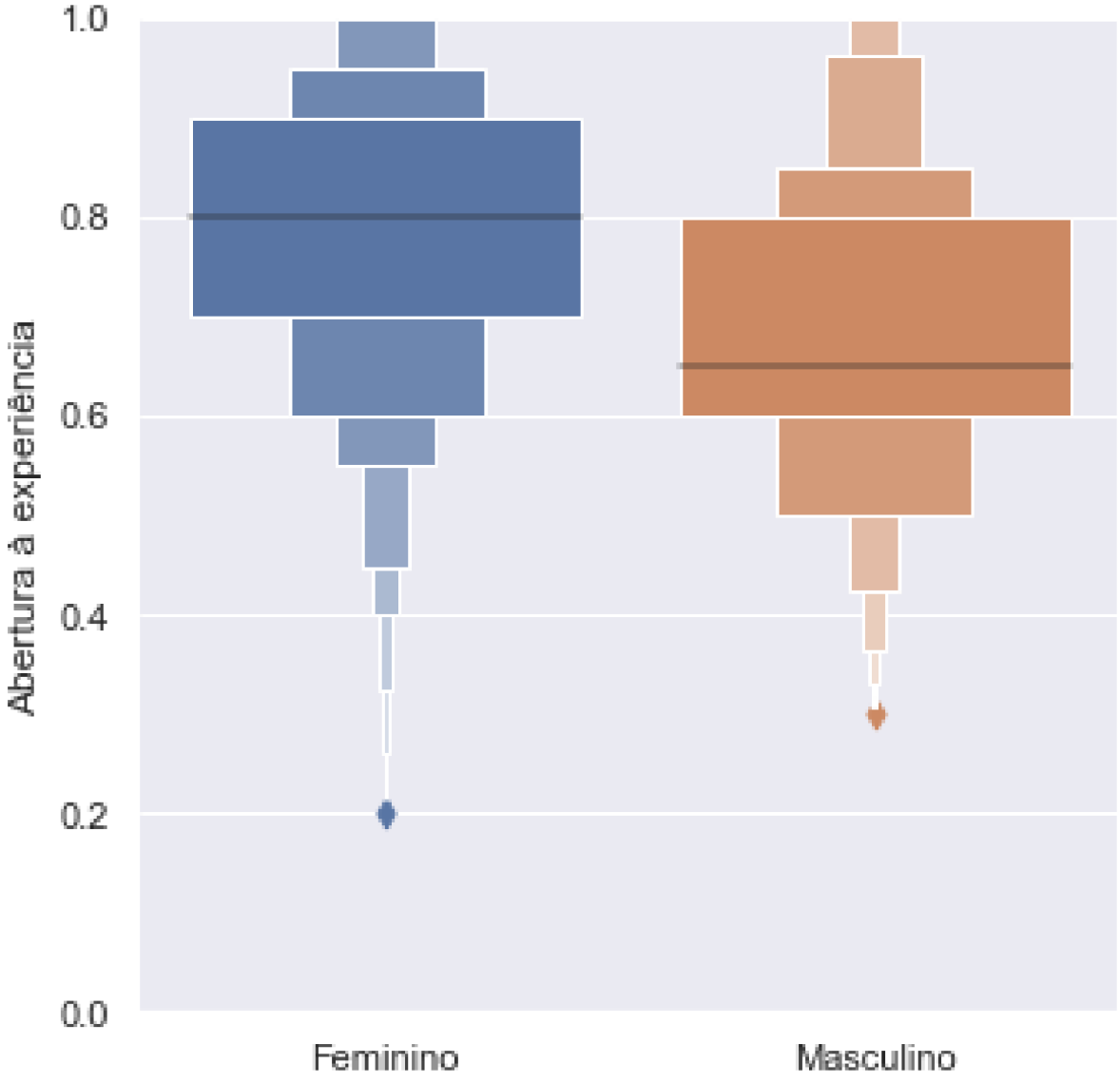

Figura 37: Exemplo de diferença no traços de personalidade de "abertura à experiência"entre os usuários por gênero

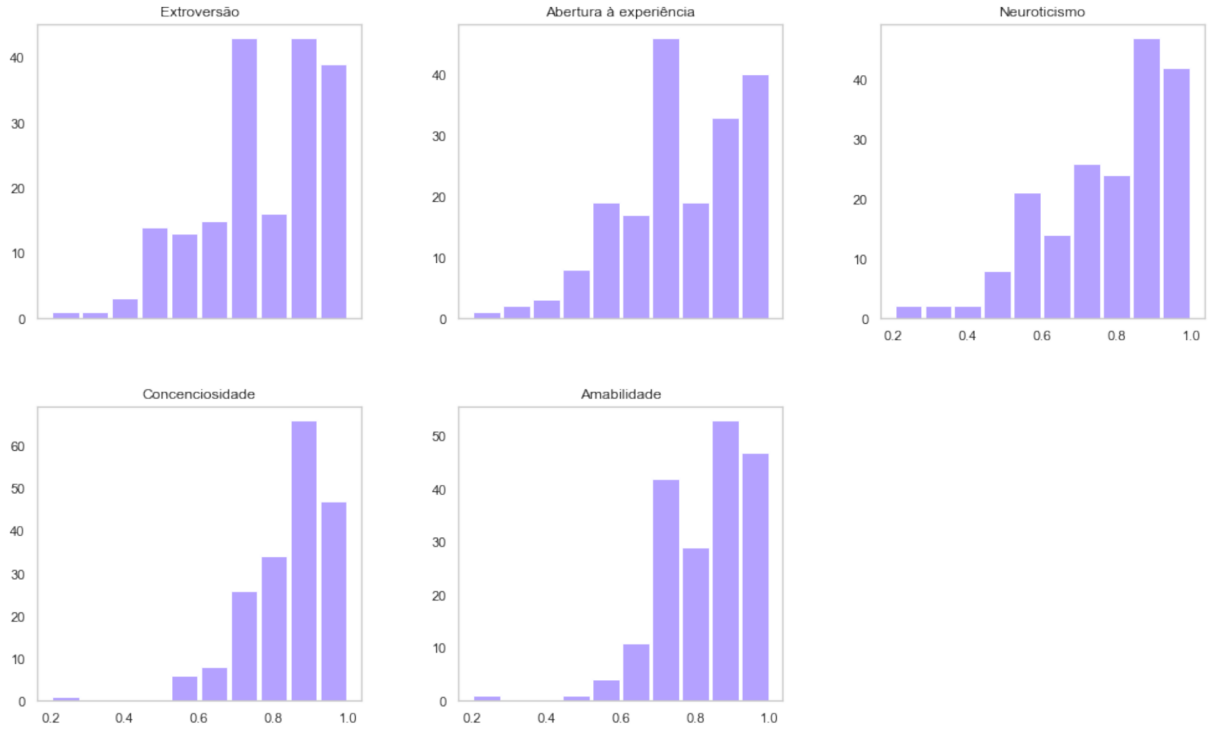

Figura 38: Histograma dos traços de personalidade no conjunto de dados combinado

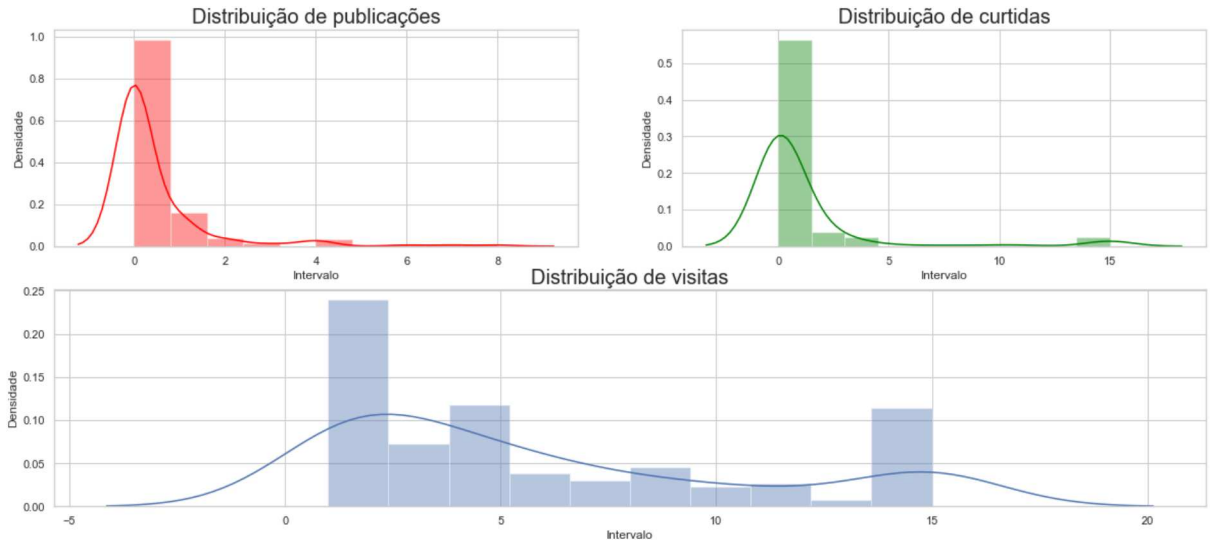

Figura 39: Distribuição comportamental nas principais atividades da rede social no conjunto de dados combinado

Tabela 7: Dimensões com maiores desvios padrões

Dimensão	Desvio Padrão	Intervalo
guests_vists	600.997806	(0, 5988)
total_visits	20.503578	(0, 164)
visits_received	12.815361	(0, 101)
event_planning_duration_months	12.564547	(0.8, 74.5)
total_likes	6.914746	(0, 66)
gifts_received	3.249023	(0, 26)
likes_received	3.099641	(0, 20)

7.3.1 Limpeza dos dados

Alguns dados do conjunto de dados podem ser definitivamente *outliers*[2], potencialmente de administradores da rede social ou de usuários que podem estar apenas explorando a ferramenta por outras razões desconhecidas ou por outros ruídos não identificados. Utilizando um corte de restrição mais brando, para remoção dos *outliers*, aplicou-se o método "The Empirical Rule", onde 99.7% dos valores de uma distribuição normal encontram-se dentro da faixa de três desvios padrão, tanto para mais quanto para menos em relação à média (JAWLIK, 2016), resultando na remoção específica dos ruídos presentes nas colunas com maiores desvios padrões do conjunto de dados de trabalho, como detalhado na Tabela 7.

Há uma pequena quantidade de valores muito grandes para algumas dimensões. Um método comum para pontuar anomalias em dados de uma dimensão é o z-score. Como a média e o desvio padrão são conhecidos, então, para cada ponto de dados, foi calculado esse índice. Portanto, foi removido do conjunto de dados todos os *outliers*, contendo valores de desvio padrão muito fortes. Esse foi o caso para as dimensões de total visitas de convidados (*guests_visits*) e total de visitas recebidas em publicações (*visits_received*). Nas demais dimensões com desvio padrão mais alto, nenhum *outlier* foi identificado e removido seguindo esse método, como pode ser observado na Tabela 8. Como resultado final dessa etapa, foram removidos ao todo quatro usuários do conjunto de dados, resultando em conjunto de dados de trabalho após essa etapa de 184 registros de usuários.

7.3.2 Imputação de valores ausentes

Nos dados disponibilizados pela rede social, há alguns casos em que elementos particulares estão ausente. Isso pode ter ocorrido devido a algumas razões, como dados corrompidos, falha em carregar as informações ou extração incompleta. Para manipular os valores ausentes e auxiliar a geração de um modelo robusto de clusterização, foi realizada uma análise dos valores

[2]Um *outlier* é definido como uma observação que “parece” ser inconsistente com outras observações no conjunto de dados. Um *outlier* tem uma baixa probabilidade de que origina-se da mesma distribuição estatística que as outras observações no conjunto de dados (WALFISH, 2006).

Tabela 8: Dimensões com maiores desvios padrões após remoção de *outliers*

Dimensão	Desvio Padrão	Intervalo
guests_vists	215.881723	(0, 1654)
total_visits	20.585464	(0, 164)
visits_received	10.731053	(0, 79)
event_planning_duration_months	12.487724	(0.8, 74.5)
total_likes	6.959576	(0, 66)
gifts_received	3.093997	(0, 26)
likes_received	2.922450	(0, 20)

ausentes para cada dimensão. Na Tabela 9, pode se observar que 94% dos dados disponibilizados não apresentam a frequência de uso mensal na rede social (*monthly_retention*) [3]. Como esse é um valor muito alto englobando a grande maioria dos exemplos, foi decidido remover essa dimensão do conjunto de dados. Para a dimensão que representa a localização regional (*state*) de cada usuário, 23% dos registros não possuem essa dimensão informada. Nesse caso específico, optou-se por transformar esse valor nulo, em um valor definido como "Não Identificado". A mesma estratégia foi aplicada para gênero (*gender*), data do casamento (*event_date*) e data de inscrição na rede social (*signed_up*).

Os tópicos de interesse de usuário apresentaram uma ausência de informações igualitárias em todo o conjunto de dados. Isso significa que 35 registros (19%) de usuários não possuem qualquer registro de interação na rede social. Analisando o conjunto de dados, verificou-se que esses mesmos usuários, de fato, não realizaram nenhuma visita ou outra interação em qualquer publicação da rede social da Wedy - potencialmente sendo usuários de outras funcionalidades do produto digital. Como o objetivo da pesquisa restringe-se a combinação dos dados de personalidade, comportamento e demografia, optou-se por seguir o mesmo método comumente usado para lidar com os valores nulos. Ou seja, foi realizada a remoção completa de dados com valores ausentes em busca de um modelo mais intencional, robusto e preciso, relativando à perda de informações sobre os traços de personalidade desses usuários.

Como resultado dessas operações de limpeza de dados e imputação de dados ausentes, o conjunto de dados chegou ao número final de 158 registros de interesse para o estudo e 446 colunas representando as dimensões originais e as novas criadas pela combinação delas e técnicas aplicadas e descritas anteriormente. Como efeito colateral da remoção desses registros, uma série de tópicos de interesse ficaram sem entradas de valores dos usuários, resultando em uma coleção de tópicos órfãos. Ao todo 95 tópicos deixaram de fazer referência a qualquer usuário, como foram os casos de "Vestido Branco", "Swarovski", "1ª Vez", "Mãe De Noiva", "Off White", "Zara"e outros. Para otimizar a matriz e evitar dimensões sem valor visível ao conjunto, essas dimensões também foram removidas, resultando uma matriz de 149 registros por 351 colunas.

[3] Os dados de retenção mensal não presentes no conjunto de dados não foram coletadas pela Wedy por falhas sistêmicas na coleta e disponibilização desses dados na granularidade necessária para essa pesquisa.

Tabela 9: Dimensões com ausência de valores

Dimensão	Usuários sem registros
monthly_retention	177
state	44
Igreja	35
Dentro De Casa	35
Criança	35
gender	9
event_date	5
signed_up	1

7.3.3 Discretização

Como pode-se observar na descrição das dimensões presentes (C.1), o conjunto de dados possui tipos mistos de dados: categóricos e discretos. Embora o método k-means seja bem conhecido por sua eficiência em agrupar grandes conjuntos de dados trabalhando apenas com dados numéricos, o mesmo não pode ser aplicado para agrupar dados categóricos (NGUYEN et al., 2019). Como revisado na Seção 4, a premissa básica do agrupamento é a similaridade ou dissimilaridade entre os pontos de dados. Ou seja, algoritmo K-means itera continuamente até atingir um estado em que todos os pontos de um cluster são semelhantes entre si e os pontos pertencentes a diferentes clusters são diferentes uns dos outros.

Analisando o conjunto de dimensões de dados categóricos, observa-se que eles representam um conjunto pequeno de informações: 1.7%. São eles: data de inscrição na rede social (*signed_up*), estado (*state*), gênero (*gender*)', data do evento (*event_date*), conteúdo textual publicado (*published_content_text*) e links das imagens públicadas (*published_content_image*)]. Para cada um dos tipos desses dados optou-se por utilizar abordagens distintas:

- **Gênero e estado:** *Label encoding* foi a abordagem escolhida. A técnica simplesmente atribui um valor inteiro a cada valor possível de uma variável categórica. Ela é uma técnica relativamente difundida e funciona particularmente bem para casos onde a variável categórica não assuma um conjunto grande de valores (HANCOCK; KHOSHGOFTAAR, 2020). Esse é o caso dessas duas dimensões presentes no conjunto de dados dessa pesquisa - principalmente, pela localização estar restrita a poucos estados, conforme ilustrado na Figura 36) e o gênero variar em 3 opções. Embora seja simples de implementar e tem tempo de execução eficiente, ela apenas informa que dois recursos são idênticos ou diferentes, pois a distância entre categorias, agora numéricas, carregam poucas informações sobre seu grau de semelhança.

- **Data do evento e data de inscrição:** a data do evento, nesse contexto de domínio, quando analisada de forma absoluta tende a não trazer informações relevantes. Porém, alguns novos atributos podem ser extraídos dela, como por exemplo a preferência pela estação do

ano e o dia da semana escolhido para a celebração. Dessa forma, a data do evento foi transformada em duas novas dimensões: estação do ano (*event_date_season*), representadas pelos valores 1 (Outono), 2 (Inverno), 3 (Primavera), 4 (Verão) e -1 (não informado) e dia da semana do evento (*event_date_weekday*) representado pelo intervalo 0-6, onde zero é domingo e 6 é sábado. A data de inscrição, seguindo a mesma estratégia, também sofre uma transformação, mas nesse caso apenas o dia da semana de cadastro é considerado (*signed_up_weekday*), potencialmente para indicar alguma preferência de cada usuário para buscar uma solução para organizar o seu evento.

- **Conteúdo textual publicado**: Como o conjunto de dados nessa etapa de pré-processamento de dados é reduzido e a rede social em estudo apresenta uma razão de 2:10 entre publicadores e consumidores de conteúdo, não há um conjunto textual extenso nessa fase da pesquisa que justifique a utilização de técnicas avançadas de processamento de dados. Por isso, realizou-se a discretização dessa dimensão em (a) uma nova dimensão (*published_content_text_sentiment*) representando uma análise simplificada de sentimento nos conteúdos publicados pelos usuários através dos valores 0 (sentimento negativo), 1 (sentimento positivo), -1 (ausência de conteúdo para inferência) e (b) novas dimensões, a partir de aplicação de LIWC básico, como número de palavras escritas (*published_content_word_count*), número de caracteres escritos (*published_content_char_count*), número de palavras que indicam processos de cognição (*published_content_cogmech_count*), número de palavras que indicam processos afetivos (*published_content_affect_count*), número de palavras que indicam processos sociais (*published_content_social_count*), número de verbos utilizados (*published_content_verb_count*) e o número de palavras que indicam negações (*published_content_negate_count*).

- **Conteúdo em imagens publicado**: a rede social Wedy já utiliza de visão computacional para extrair os tópicos inerentes a cada imagem publicada - representada aqui nesse conjunto de dados nos tópicos de interesse de cada usuário. Portanto, para simplificações no modelo de clusterização, os conteúdos visuais não serão pré-processados e utilizados apenas a nível de estudo dos agrupamentos gerados. Assim, a técnica utilizada para discretizar essa coluna é a transformação simples para um contador de imagens publicadas para cada usuário, originando uma nova dimensão *published_content_image_count*.

A partir do conjunto de respostas do questionário de inferência de personalidade, como observado na Seção 6.4.2, é possível entender as diferenças entre os tipos de personalidade dos usuários da rede social em um nível mais amplo. Para facilitar o agrupamento dos cinco traços de personalidades descritos pelo modelo utilizado, os valores foram distribuídos em uma série de medidas. Para cada medida, as pontuações foram divididas em quartis dinâmicos para produzir medidas dessa característica que podem variar de baixa a alta, uma técnica utilizada por diversos trabalhos na área (WHAITE et al., 2018) (ADLAI-GAIL, 1995) (TERRACCIANO et al., 2014). Dessa forma, o quartil 1 foi classificado como um fator de medida mais baixa, o

quartil 2 como um fator de medida baixa, o quartil 3 como um fator de medida alta e o quartil 4 como um fator de medida mais alta. Por exemplo, um usuário que possui um índice de 0,51 no traço de personalidade amabilidade, possui uma medida baixa nesse fator (quartil 2). Por outro lado, um usuário com um índice de 0,97 em "neuroticismo", possui uma medida mais alta nesse fator (quartil 4), dentro do contexto observado na rede social Wedy. Os quartis podem facilitar análises futuras, como potencialmente indicar que um usuário com alto neuroticismo e baixa amabilidade é mais ativo na rede social, e produz conteúdos mais negativos em relação a média de usuários da rede social. A distribuição do conjunto de dados agrupado pelos cinco traços de personalidade, agrupados pelos quatro quartis, fornece 1024 combinações (4^5), e pode ser observado na Tabela 10.

Para criar esses quatro grupos de representação de valores de cada traço de personalidade, escolheu-se discretizar as variáveis referentes em intervalos de tamanhos iguais com base em quantis de amostra. Isso basicamente significa que o conjunto de valores de cada variável é dividido em dados subjacentes em compartimentos de tamanhos iguais. A função utilizada "*Quantile-based discretization*"define os intervalos usando percentis com base na distribuição dos dados.

Com a discretização de todas variáveis categóricas, finaliza-se essa etapa com a transformação de algumas dimensões e a criação de novas dimensões a partir delas. A Figura 40 demonstra a distribuição dessas dimensões que passam a incorporar o conjunto de dados de trabalho, que agora passa de 351 colunas para 360 colunas, mantendo o mesmo número de linhas (registros de usuários), permitindo que a pesquisa avance para a próxima etapa de pré-processamento de dados, o escalonamento dos dados discretos.

Traço de personalidade	Amostra Total	Amostra Percentual	Intervalo	Média
Neuroticismo	**149**	**100%**	**0,20-1.00**	**0,78**
Mais baixa	*38*	*26%*	*0,20-0,65*	*0,54*
Baixa	*42*	*28%*	*0.70-0,80*	*0.76*
Alta	*37*	*25%*	*0,85-0,90*	*0.87*
Mais alta	*32*	*21%*	*0,95-1,00*	*0.98*
Extroversão	**149**	**100%**	**0,35-1,00**	**0,78**
Mais baixa	*55*	*37%*	*0,35-0,70*	*0,60*
Baixa	*27*	*18%*	*0,75-0,75*	*0,77*
Alta	*36*	*24%*	*0,85-0,90*	*0,87*
Mais alta	*31*	*20%*	*0,95-1,00*	*0,97*
Abertura à experiência	**149**	**100%**	**0,30-1,00**	**0,76**
Mais baixa	*42*	*28%*	*0,30-0,65*	*0,56*
Baixa	*38*	*26%*	*0,70-0,75*	*0,72*
Alta	*37*	*25%*	*0,80-0,85*	*0,84*
Mais alta	*32*	*21%*	*0,95-1,00*	*0,97*
Concensiosidade	**149**	**100%**	**0,55-1,00**	**0,85**
Mais baixa	*55*	*37%*	*0,55-0,80*	*0,74*
Baixa	*28*	*19%*	*0,85-,0,85*	*0,85*
Alta	*39*	*26%*	*0,90-0,95*	*0,91*
Mais alta	*27*	*18%*	*1,00-1,00*	*1,00*
Amabilidade	**149**	**100%**	**0,50-1,00**	**0,83**
Mais baixa	*47*	*32%*	*0,50-0,75*	*0,69*
Baixa	*43*	*29%*	*0,80-0,85*	*0,82*
Alta	*38*	*26%*	*0,90-0,95*	*0,92*
Mais alta	*21*	*14%*	*1,00-1,00*	*1,00*

Tabela 10: Traços de personalidade agrupados em quartis por fator de intensidade no contexto do conjunto de dados

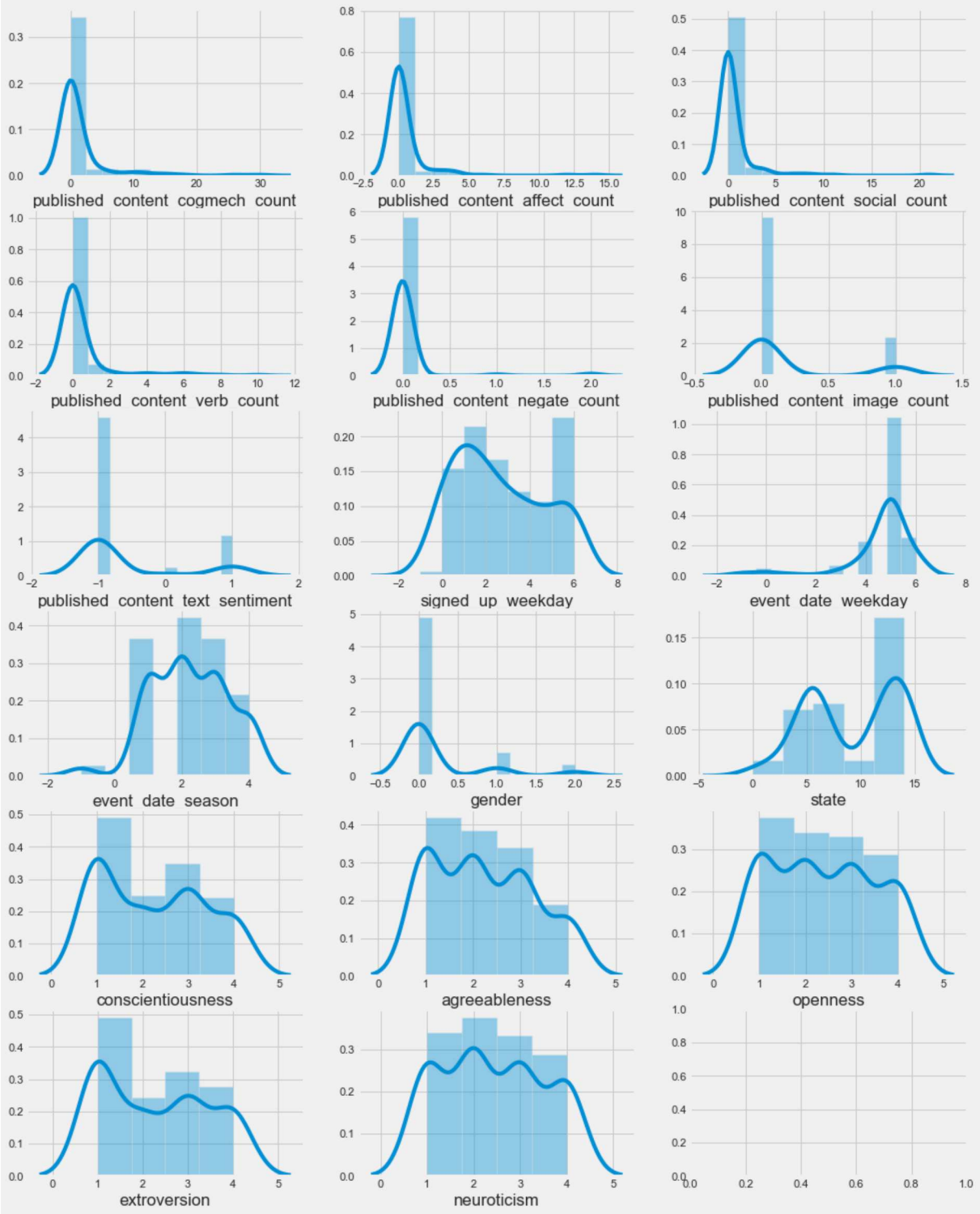

Figura 40: Novas dimensões e dimensões transformadas após a aplicação de diferentes técnicas de discretização

7.3.4 Escalonamento dos dados

Após a limpeza inicial de dados, pode-se observar que cada dimensão do conjunto de dados, embora discretizadas, tem significados diferentes e, obviamente, apresentam uma distribuição diferente de valores. Portanto, para trabalhar com uma distribuição semelhante (com valores esperados iguais a 1), o escalonamento desses valores é uma etapa importante para evitar que o algoritmo de clusterização fique enviesado para as variáveis com maior ordem de grandeza. O agrupamento de K-means, algoritmo candidato nas próximas etapas do trabalho, é isotrópico em todas as direções do espaço e, portanto, tende a produzir agrupamentos mais ou menos redondos (em vez de alongados). Nessa situação, deixar variâncias desiguais equivale a colocar mais peso em variáveis com variância menor, de modo que os clusters tendem a ser separados ao longo de variáveis com variância maior. Ao dimensionar as variáveis, a comparação de diferentes variáveis ocorre em pé de igualdade. Ou seja, ao escalonar e padronizar os dados brutos, convertendo-os em intervalo específico usando uma transformação linear, pode-se potencialmente gerar clusters de melhor qualidade e melhorar a precisão dos algoritmos de agrupamento.

Para essa etapa de escalonamento (ou padronização) dos dados, optou-se por remover a coluna de identificador do usuário do conjunto de dados, pois cada registro do conjunto já representa as informações de um usuário. Além disso, a coluna apresenta números randômicos de alta grandeza, tornando-se, além de irrelevante, um ruído extra para o algoritmo de escalonamento. O algoritmo escolhido foi o *Min-max*, que baseia-se no processo de capturar os dados mensurados em suas unidades singulares e transforma-los em um valor entre 0,0 e 1,0, onde o valor mais baixo (*valor mínimo*) é definido como 0,0 e o mais alto (*valor máximo*) como 1,0. Isso fornece uma maneira fácil de comparação dos valores mensurados usando escalas diferentes ou unidades de medida diferentes (MOHAMAD; USMAN, 2013). O resultado está disposto na Figura 41, demonstrando as curvas simétricas individuais de cada uma das 358 dimensões. Esta, também, é uma técnica para garantir que todas as variáveis numéricas estejam aproximadamente na mesma escala que as categóricas (binárias). O algoritmo *Min-max* também fornece uma função de escalonamento inverso para recuperação dos valores originais, fundamental para análise qualitativa e validação das hipóteses propostas pelo trabalho.

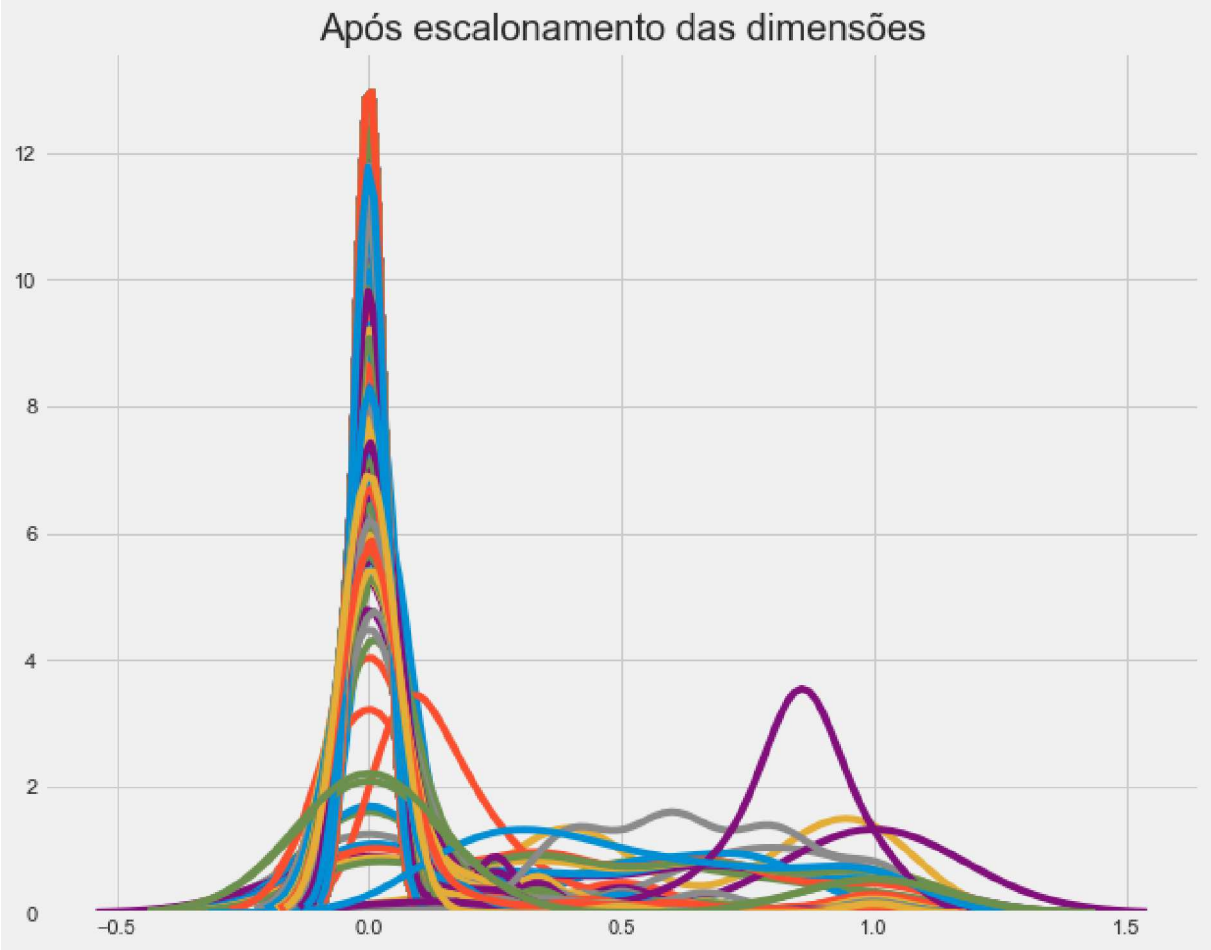

Figura 41: Resultado da aplicação de escalonamento para cada caso de assimetria transformando a distribuição de cada uma delas em normal.

7.4 Clusterização aplicada

A clusterização, como revisado na Seção 4, é um método de aprendizado de máquina não supervisionado para particionar o conjunto de dados em um conjunto de grupos ou clusters. Um desafio dessa abordagem é que os métodos de clusterização retornam grupos, mesmo que esses conjuntos não sejam relevantes. Portanto, é necessário avaliar a tendência de agrupamento, antes da análise dos grupos gerados, para validar a qualidade do resultado após a formação dos agrupamentos, tornando a validação de agrupamento importante para projetar o procedimento de avaliação dos resultados de um algoritmo de cluster.

Dada a variedade de medidas propostas na literatura para avaliar os resultados de técnicas de clusterização, bem como os principais indicadores escolhidos para essa pesquisa e apresentados na Seção 6.5.3.1, optou-se por realizar uma validação relativa de cluster, avaliando a estrutura do cluster variando valores de parâmetros (por exemplo: variando o número de clusters k), para diferentes algoritmos e utilizando conjunto de características selecionadas distintas.

7.4.1 Algoritmos selecionados

Embora o **K-means** seja o algoritmo mais simples e comumente usado - e inicialmente tenha se iniciada a exploração por ele, outros três algoritmos foram comparados com ele para a validação relativa dos agrupamentos formados

- **Spectral Clustering**: ao contrário de k-means, onde os clusters resultantes formam conjuntos convexos, o agrupamento espectral pode resolver problemas gerais como espirais entrelaçadas e trabalhar de forma eficiente mesmo para conjuntos de dados esparsos (VON LUXBURG, 2007), e foi selecionado para ser um contraponto importante aos demais;

- **Agglomerative Clustering**: escolhido como uma opção especialista de algoritmos hierárquico de clusterização, é uma abordagem "de baixo para cima", onde cada observação começa em seu próprio agrupamento e pares de agrupamentos são mesclados à medida que se sobe na hierarquia (MÜLLNER, 2011).

- **K-medoides**: Da mesma classe de algoritmos que o K-means, o algoritmo K-Medoids é usado para encontrar *medoids*[4] em um cluster o qual está localizado em um ponto central. Foi escolhido, por ser mais robusto em comparação com K-means, com técnicas que reduzem ruídos nos dados a partir de *outliers* (ARORA; VARSHNEY et al., 2016).

[4] objetos representativos de um conjunto de dados ou um cluster dentro de um conjunto de dados cuja dissimilaridade média para todos os objetos no cluster é mínima

7.4.2 Seleção de características exploradas

Como este trabalho tem como objetivo explorar um conjunto de dados esparso, em uma combinação de vetores distintos de dados, optou-se, mesmo com uma quantidade de registros pequenos a serem agrupados devido às questões exploradas na Seção 6.4.1, em manter um conjunto de dados com todas as dimensões, e explorar de forma adjacente outros dois conjuntos de dados representando as principais dimensões definidas por técnicas de redução de dimensionalidade e outro conjunto representando apenas os traços de personalidade e os índices sintéticos de comportamento disponibilizados para o estudo. As três estratégias de seleção de características são apresentadas a seguir:

- **Todas Dimensões**: seleção que compreende todas as características mantidas após o processo de pré-processamento inicial descritas na Seção 7.3. Dessa forma, esse é o conjunto que representa uma maior esparsidade da matriz, ou seja, a seleção de características que representa maior heterogeneidade dos dados;

- **Principais Dimensões**: seleção definida a partir de técnicas de redução de dimensionalidade, onde características constantes ou com variância mínima, os quais não fornecem informações que permitam a um modelo de aprendizado de máquina diferenciar potenciais grupos a partir deles. Para identificar essas características, foi utilizada a técnica de *Variance Threshold*. Nesse método, as características são removidas tal qual a sua variância não exceda um certo valor de limite, ou seja, se uma característica está carregando informação similar em todo conjunto, ela é pouco relevante para a geração de agrupamentos significativos e pode ser removida (CHAUHAN; MATHEWS, 2019). Na Figura 42 é possível verificar o modelo utilizado que definiu o limite em 0.8 e 97 características a partir dele, o qual representa visualmente o limítrofe para a aplicação da técnica.

- **Traços de Personalidade & Principais índices de comportamento**: seleção definida a partir do volume de dados úteis do conjunto de dados finais (116 registros), representada pelos cinco traços de personalidade e pela simplificação da seleção de características apenas pelos totalizadores de ações dos conjuntos comportamentais, tais como (total de visitas, curtidas, comentários, publicações e outros similares). Os dados demográficos foram descartados nesse conjunto.

7.4.3 Número de agrupamentos validados

Determinar o número ideal de clusters em um conjunto de dados é a questão fundamental no agrupamento de dados via técnicas de clusterização. Todos os algoritmos selecionados aqui para essa pesquisa, requerem a especificação do número de clusters k a ser gerado. Para determinar esse número ideal de clusters, optou-se pelo intervalo mínimo de dois grupos e o

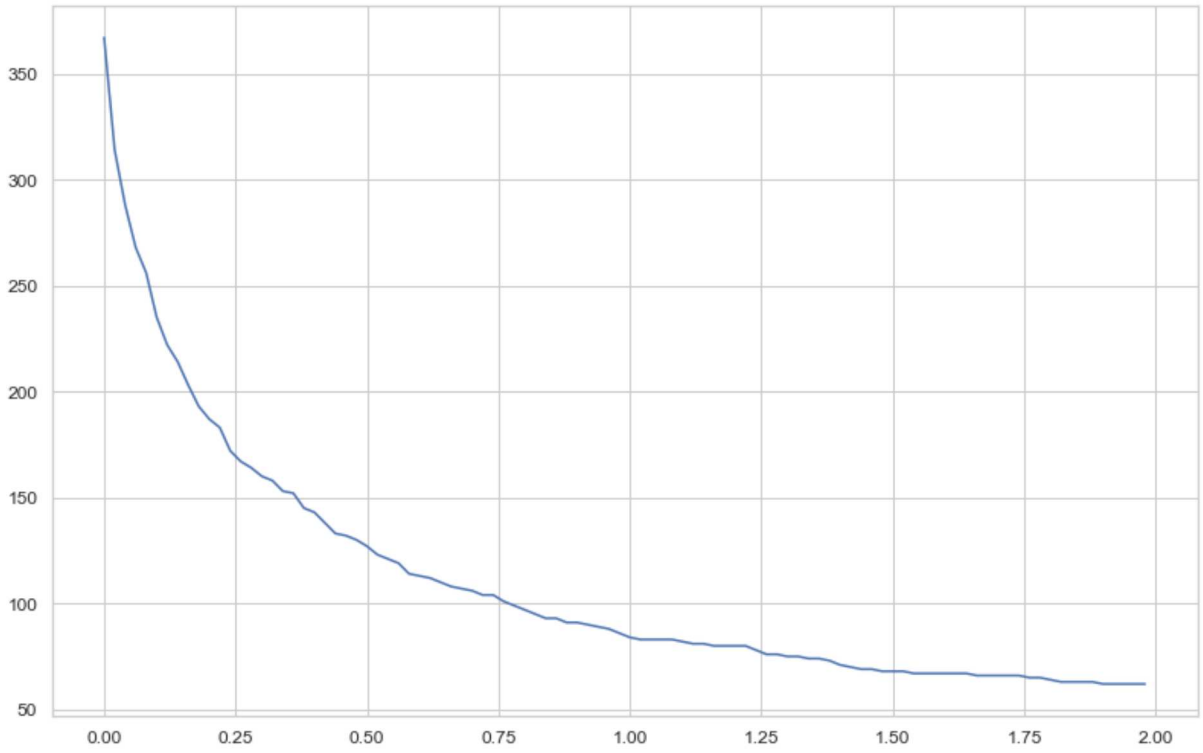

Figura 42: Curva de limite para escolha de características com maior variância entre o conjunto de dados

máximo de 10 grupos, definido pelo tamanho do conjunto de dados, a fim de não gerar grupos extremamente fragmentados. A partir desse intervalo, os três conjuntos de dados e os quatro algoritmos previamente selecionados foram aplicados de forma incremental a fim de se analisar as principais métricas destacadas na Seção 6.5.3.1.

7.4.4 Análise descritiva dos agrupamentos gerados

Para estudar o modelo de agrupamento com os melhores índices, todas as combinações entre número de clusters esperados, algoritmos previamente selecionados e diferentes seleções de recursos foram aplicadas. Os resultados de cada uma dessas aplicações estão representados na Figura 43, com os seus devidos índices de DBI (a pontuação mínima é zero, com valores mais baixos indicando melhor agrupamento) e SC (com o melhor valor sendo 1 e o pior valor sendo -1; valores próximos a 0 indicando clusters sobrepostos; e valores negativos geralmente indicando que uma amostra foi atribuída ao cluster errado).

Os clusters extraídos dos conjuntos de dados são baseados em uma medida de distância ou correlação. Isso resulta em grupos coerentes, onde os indivíduos do mesmo grupo têm características semelhantes. Os dois índices de validade de cluster, ou seja, DBI e SC, fornecem a medida de quão bem separados esses clusters estão uns dos outros, além da coesão interna do cluster.

Como processo dessas validações, 96 estratégias de agrupamentos distintos foram utilizadas

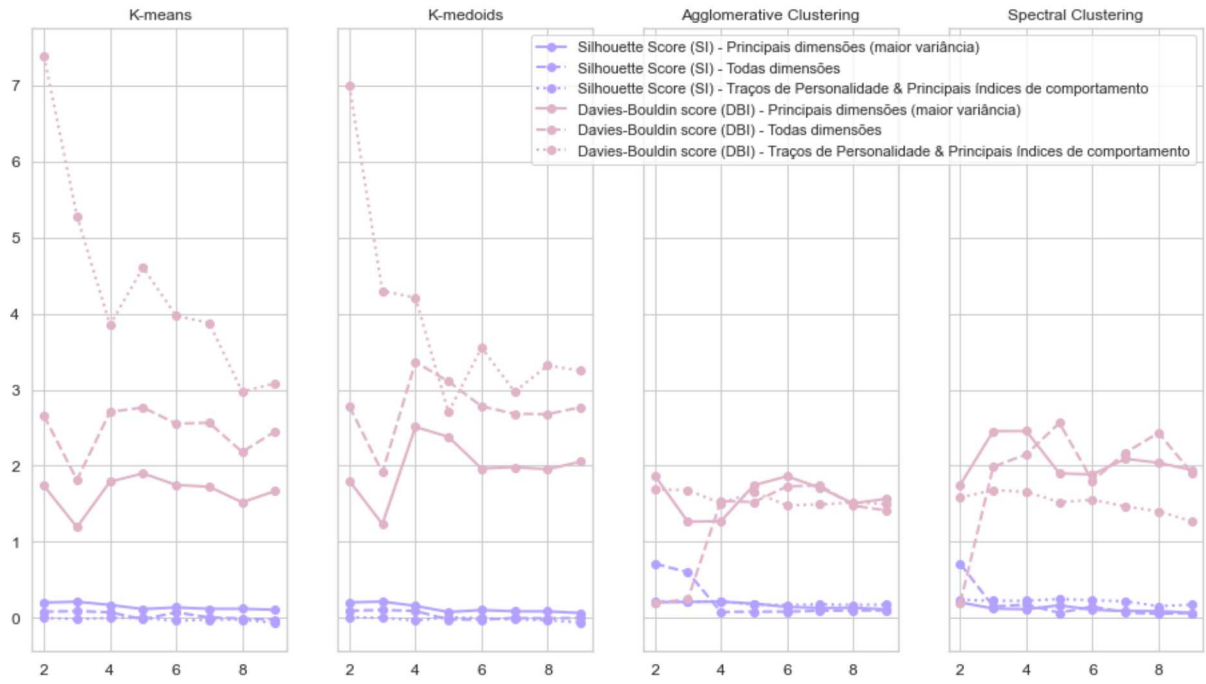

Figura 43: Os valores DBI e SC para as várias formações de agrupamento do conjunto de dados usando K-means, K-medoides, Agglomerative Clustering e Spectral Clustering

(com quatro algoritmos de clusterização distintos, oito opções de segmentações de número de clusters e três conjuntos de seleção de características distintos). Ou seja, para cada conjunto de características, foi inicialmente selecionada a melhor seleção de clusters (k) configurada em cada algoritmo. A partir do desempenho de cada algoritmo em seus melhores números de clusters, observados pelos índices de SC e DBI, selecionou-se o melhor modelo para cada conjunto de características. Consequentemente, observou-se os seguintes resultados:

- **Todas Dimensões**: Spectral Clustering com SC de 0.222 e DBI de 1.938 com quatro grupos ($k = 4$). Ressalta-se que em todas combinações de algoritmos e número de clusters, utilizando-se de todas as dimensões, sempre foram percebidos clusters desbalanceados, com grupos formados por apenas um único elemento;

- **Principais Dimensões**: K-Means com SC de 0.219 e DBI de 1.211 em três grupos ($k = 3$). Nesse caso também foi encontrado um agrupamento com apenas um único elemento, potencialmente representando um *outlier* para o modelo;

- **Traços de Personalidade & Principais Índices de Comportamento**: Spectral Clustering com SC de 0.241 e DBI de 1.616 em dois grupos ($k = 2$). Ressalta-se que entre todos conjuntos de características, esse foi o único a apresentar um resultado balanceado, com agrupamentos bem distribuídos.

Como descrito anteriormente e observado na Tabela 11, o conjunto de características denominado Principais Dimensões, possui a melhor dupla combinada de índices SC e DBI, quando o

	Todas Dimensões				Principais Dimensões				Traços de Personalidade & Índices de Comportamento			
Índices	**K-Means**	**K-Medoid**	**Agl. Clustering**	**Spectral Clustering**	**K-Means**	**K-Medoid**	**Agl. Clustering**	**Spectral Clustering**	**K-Means**	**K-Medoid**	**Agl. Clustering**	**Spectral Clustering**
SC	0.103	0.094	0.085	**0.222**	**0.219**	0.212	0.207	0.212	0.008	-0.054	0.212	**0.241**
DBI	1.822	2.691	1.903	**1.938**	**1.211**	1.757	1.193	1.757	2.424	3.604	1.738	**1.616**
Clusters (k)	3	2	4	**4**	**3**	2	3	2	2	6	2	**2**

Tabela 11: Comparação dos melhores índices em diferentes técnicas de clusterização para os três conjuntos de características

algoritmo selecionado é o K-Means com três clusters ($k = 3$). Como um dos agrupamentos gerados por esse modelo possui apenas um único elemento, optou-se por tratá-lo como um *outlier* (uma contra todas as amostras), descartando esse agrupamento. Dessa forma, esse foi o modelo escolhido para ser trabalhado. Secundariamente, o conjunto de "Traços de Personalidade & Principais índices de comportamento", por representar índices relevantes no comparativo com os demais, por ter uma redução significativa na esparsidade dos dados e estar com agrupamentos distribuídos de forma balanceada, também é candidato a ser observado e utilizado como modelo comparativo.

Para estudar as propriedades do cluster, é necessária prosseguir para uma análise descritiva. Normalmente, a análise descritiva é baseada no conjunto de características utilizadas no processo de clusterização. Nesta proposta, como descrito anteriormente, optou-se por detalhar o conjunto com as Principais Dimensões e utilizar como comparativo o conjunto "Traços de Personalidade & Principais Índices de Comportamento". Ambos os conjuntos contém dimensões de características socioafetivas, bem como o comportamento de usuário registrados por pegadas digitais ativas e passivas.

Utilizando uma abordagem supervisionada, pode-se avaliar a contribuição das variáveis presentes em cada um dos conjuntos de dados na formação dos agrupamento (ISMAILI; LEMAIRE; CORNUÉJOLS, 2014). Dessa forma, é possível compreender um agrupamento de dados de alta dimensão, o que é relevante ao analisar o conjunto Principais Dimensões que possui 85 dimensões. Utilizando-se de uma estratégia supervisionada de classificação multi-classe e tendo como alvo o cluster gerado (isto é, Cluster A ou Cluster B e considerando os elementos que pertencem a cada cluster membros da mesma classe), treinou-se um classificador a partir do Random Forest. Esse é um algoritmo clássico que utiliza-se de uma combinação de preditores em árvores, de modo que cada árvore depende dos valores de um vetor aleatório de forma independente e com a mesma distribuição para todas as árvores disponíveis. O algoritmo obteve uma acurácia de 90% quando otimizado para 10 árvores de decisão. Na Figura 44, pode-se observar os coeficientes das variáveis do classificador, que podem servir para estimar a importância

de cada variável nos agrupamentos de usuários utilizando-se do conjunto de características das Principais Dimensões. Isto posto, analisando as dimensões mais relevantes, nota-se que:

- **Pegadas digitais passivas:** A segunda variável mais relevante na formação dos clusters trata-se de uma pegada digital passiva, o total de visitas recebidas nos conteúdos publicados, as quais não deixam rastros visíveis na plataforma. Além disso, das 40 dimensões com coeficiente de contribuição maior, 25 são pegadas digitais passivas, como o tipo de conteúdo consumido, entre eles Vestido Longo, Amor, Feminino, Convite, Deus, Sapato, Rosto, Roupas, Amor e outros;

- **Análise de componentes linguísticos e emocionais**: entre as 10 principais dimensões, 6 delas são referentes a análise e transformação do conteúdo textual em categorias derivadas de gramática e psicologia, bem como análise sentimental. Essas são pegadas digitais que não são caracterizadas diretamente pelos usuários, embora passíveis de exploração pelas redes sociais de forma indireta;

- **Dados demográficos**: apesar de menos relevantes que os dados comportamentais sintéticos de redes sociais como o total de publicações, o alcance de suas publicações, número de interações diversas e outros, alguns dados demográficos se mostraram relevantes na geração dos agrupamentos como: o tempo de duração do planejamento e organização do evento, o estado de origem do usuário, o dia da semana que o usuário optou por cadastrar-se na rede social e a estação do ano que optou-se pela realização do evento;

- **Traços de personalidade**: apenas dois traços de personalidade se mostraram mais relevante no contexto do conjunto de dados das Principais Dimensões. Abertura a Experiências e Neuroticismo foram as que tiveram maior destaque entre os cinco traços possíveis (Figura 45).

Seguindo a análise descritiva, baseada na formação dos dois agrupamentos criados a partir do conjunto de dados denominado Principais Dimensões, pode-se observar nas Figuras 46, 47, 48, 49 e 50 a distribuição dos valores escalonados entre os clusters, bem como a distribuição dos valores não escalonados no Anexo C.2 – com o intervalo presente para cada característica e seu valor médio. No Cluster 1 estão concentrados os usuários que realizaram mais visitas em conteúdos publicados na rede social (mais de 2x em relação ao Cluster 2). Em compensação, embora os usuários de ambos os clusters tenham um valor médio muito similar em criação de conteúdo, os usuários do Cluster 2 praticaram 4x a ação de deixar explícita o gosto (ação curtir da rede social) por um conteúdo publicado por outro usuário e de forma similar receberam 1.2x mais visitas em seus conteúdos publicados. Esse comportamento pode estar relacionado ao traço de personalidade de conscienciosidade, que como revisado na Seção 2.3, está relacionado ao ato de se expressar com cuidado e de forma minuciosa, traço de personalidade o qual é levemente superior no Cluster 1. Quando analisa-se os dados demográficos do evento, observa-se que o algoritmo optou por distribuir no Cluster 1 os índices com maior valor para

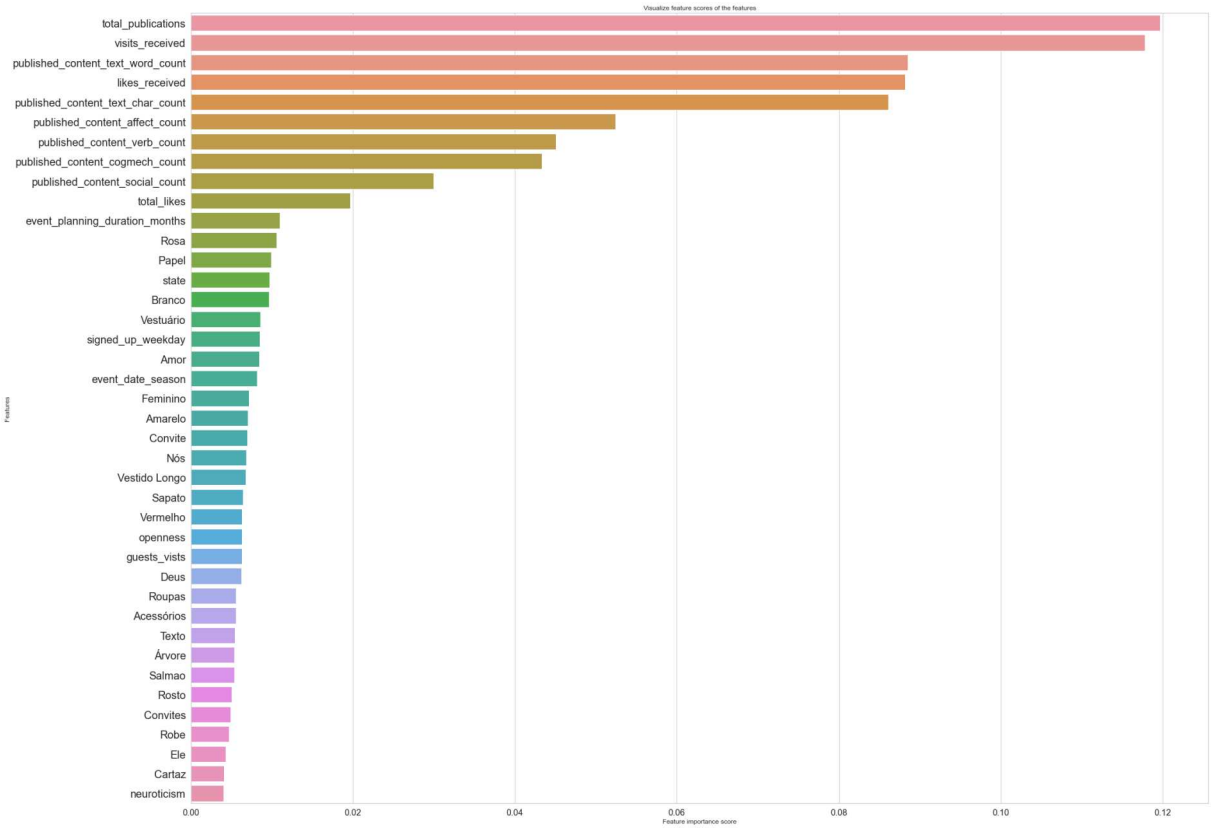

Figura 44: Coeficiente de importância de cada dimensão na formação dos agrupamentos para as Principais Dimensões

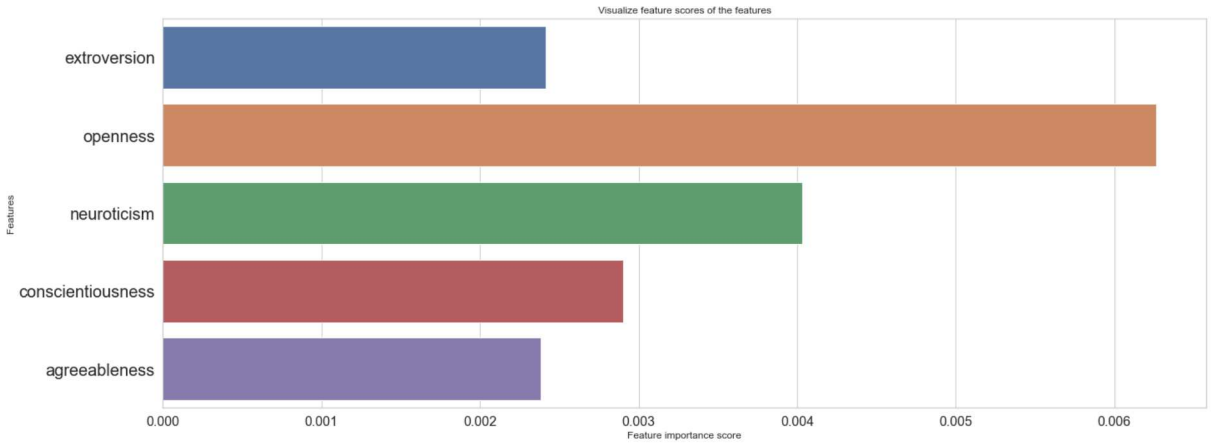

Figura 45: Coeficiente de importância dos traços de personalidade na formação dos agrupamentos para as Principais Dimensões

eventos clássicos, na cidade, e tamanhos entre pequeno/médio, ao contrário do Cluster 2 que concentrou usuários com indefinição nessas opções e, consequentemente, menor grau de planejamento, o que também pode estar relacionado ao índice mais baixo de conscienciosidade. O mesmo comportamento e cuidado com o planejamento do evento está relacionado ao tempo prévio de organização do evento, com os usuários com maior conscienciosidade descobrindo e planejamento seus eventos com 15 meses de antecedência (Cluster 1) contra 8 meses dos demais

(Cluster 2).

Embora pode-se correlacionar alguns traços demográficos e de comportamento ao traço de personalidade de conscienciosidade, os agrupamentos gerados no conjunto de dados Principais Dimensões, mostrou uma distribuição muito similar entre os demais traços - especialmente, o valor médio de abertura a experiência (2,31 contra 2,26). De todo modo, um padrão interessante que vale notar é a concentração de índices levemente superiores de extroversão e amabilidade no Cluster 1 ao mesmo tempo que esse mesmo grupo possui um menor índice de neuroticismo. Também vale destacar no Cluster 1 a potencial relação dos traços de personalidades com a distribuição comportamental e demográfica desse grupo, como (a) o maior engajamento do Cluster 1 na rede social, onde todas as categorias de conteúdo foram mais acessadas pelos usuários desse cluster, (b) o maior número médio de presentes recebidos no Cluster 1 (1.75x), (c) o maior número médio de visitas na página do evento por convidados no Cluster 1 (1.9x) e (d) a estação mais comum dos eventos no Cluster 2 foi o inverno.

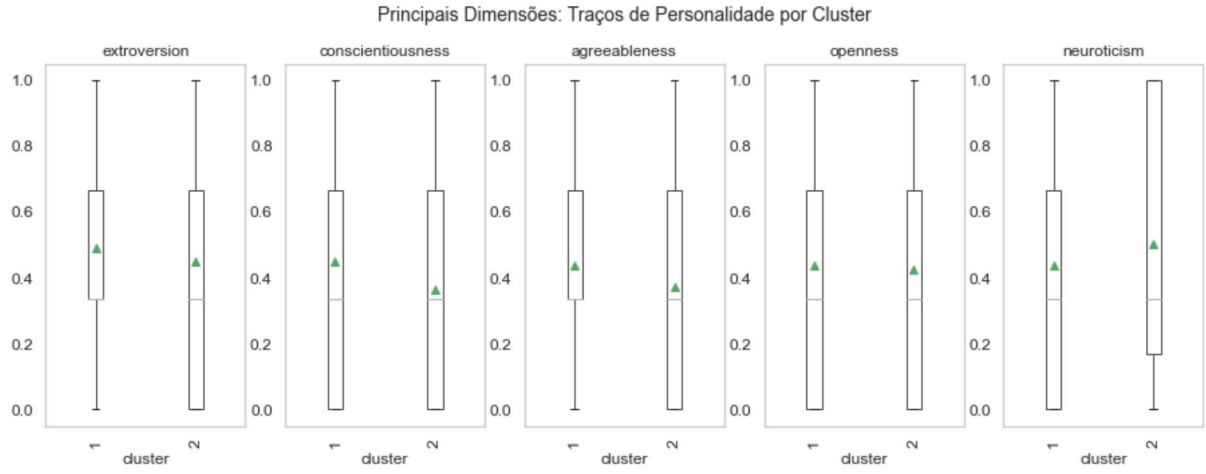

Figura 46: Análise de distribuição dos clusters no conjunto Principais Dimensões: Traços de Personalidade

De forma sintética, agrupou-se os indicadores que foram mais significativos em cada agrupamento. Na Tabela 12, é possível perceber que o Cluster 1 possui um engajamento maior na rede social, um planejamento detalhista e índices de contenciosidade, extroversão e amabilidade maiores em relação ao Cluster 2, que demonstrou um menor engajamento na rede social, bem como uma preocupação menor no planejamento dos seus eventos. Vale ressaltar que as demais dimensões foram distribuídas de forma balanceada entre os dois clusters, tais como a análise linguística (LIWC) e sentimental, bem como o estado (região).

	Cluster 1	**Cluster 2**
Engajamento	Alto	Baixo
Planejamento	Detalhista	Improvisado
Conscienciosidade	Levemente maior	Levemente menor
Neuroticismo	Levemente menor	Levemente maior

	Cluster 1	**Cluster 2**
Extroversão	Levemente maior	Levemente menor
Amabilidade	Levemente maior	Levemente menor

Tabela 12: Representação sintética das características de cada Cluster.

Como revisado anteriormente, esse conjunto de dados de Principais Dimensões, mesmo sendo uma versão reduzida do conjunto original, ainda é representado por uma matriz esparsa com dados de 157 usuários distribuidas em 86 colunas. Embora matrizes esparsas sejam comuns no aprendizado de máquina aplicado, com um conjunto de dados pequeno como o trabalhado aqui, elas tendem a apresentar muitos valores zero, sendo distintas das matrizes densas. Dessa forma, com tantas características para tão poucas observações, os modelos tendem a ajustar o ruído nos dados de treinamento, criando clusters pouco significativos, afetando negativamente o potencial de análise desejado nesse trabalho.

Como contrapartida, a análise descritiva segue a partir de agora uma observação sobre os agrupamentos formados quando utilizado apenas os principais dados de comportamento e traços de personalidade, no conjunto de dados anteriormente denominado como Traços de Personalidade Principais índices de comportamento, contabilizando apenas 12 dimensões ao todo, sete comportamentais e cinco socioafetivas. Nessa distribuição, o Cluster 1 foi formado por 92 usuários e o Cluster 2 por 65 usuários, como uma distribuição significativa nos traços de personalidade

A Figura 52 demonstra a distribuição interna dos traços de personalidade no Cluster 1. Nesse agrupamento, destaca-se uma maior pontuação em neuroticismo em relação aos demais traços que apresentam em sua totalidade valores inferiores, representando, dessa forma, na média, valores mais baixos para abertura à experiência, extroversão, conscienciosidade e amabilidade. Por outro lado, o Cluster 2, representado na Figura 52, agrupa os usuários que possuem índices mais altos de abertura à experiência, amabilidade, conscienciosidade e, principalmente, extroversão, com uma pontuação relativa inferior de neuroticismo.

Analisando a distribuição das dimensões nos dois clusters, observa-se tanto visualmente na Figura 53 com distribuição dos valores escalonados, e no Anexo C.3 a partir da distribuição dos valores não escalonados – com o intervalo presente para cada característica e seu valor médio, as características especificadas de cada agrupamento. No Cluster 2 os índices de comportamentais de engajamentos na rede social são quase na totalidade superiores aos usuários do Cluster 1, com destaque para o total de comentários deixados (2,5x superior), o total de comentários recebidos (9x superior) e o total de visitas realizadas (1.3x superior). De forma correlacionada, o Cluster 2 também apresenta os traços de personalidade esperados para esse perfil comportamental, como extroversão no índice mais alto (3,3), alta amabilidade, alta abertura à experiência (3,1), conscienciosidade alta (3,0), e neuroticismo em um valor médio (2,6). Embora o neuroti-

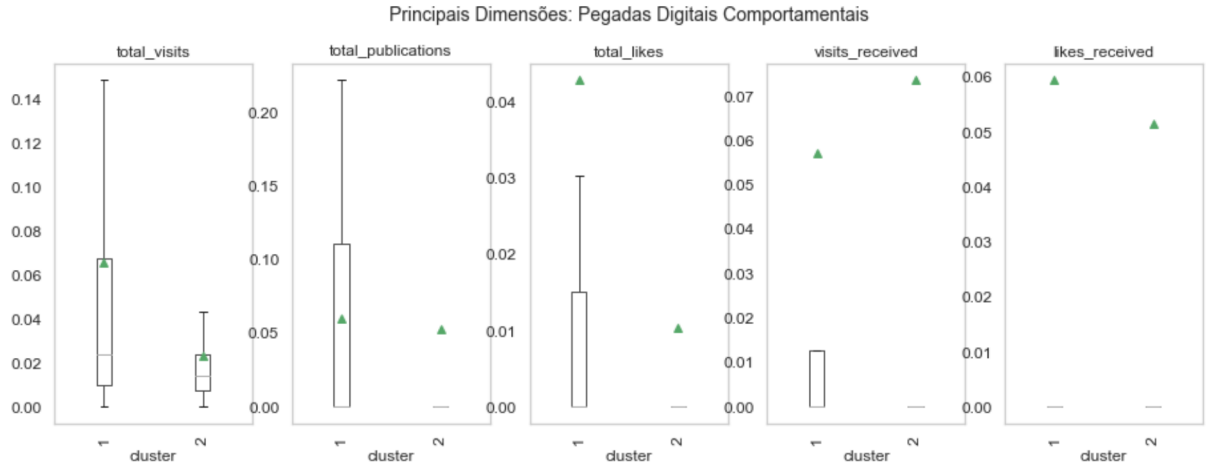

Figura 47: Análise de distribuição dos clusters no conjunto Principais Dimensões: Pegadas Digitais Comportamentais

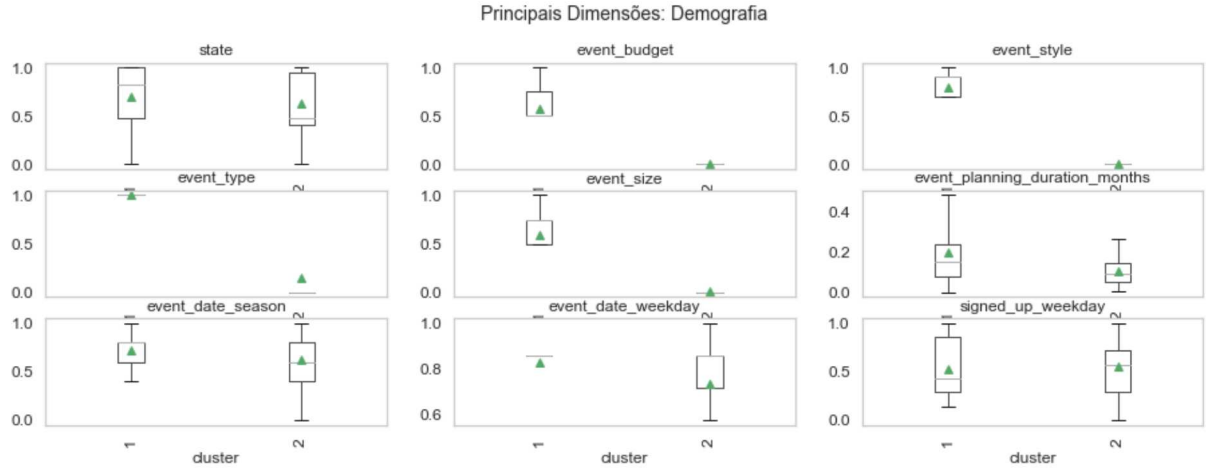

Figura 48: Análise de distribuição dos clusters no conjunto Principais Dimensões: Demografia

cismo do Cluster 1 (2,3) seja mais baixo na média que o do Cluster 2 (2,6), no auto-relato dos usuários do Cluster 1, o neuroticismo possui um valor relativo mais alto em comparação com os demais traços, onde conscienciosidade e abertura à experiência tem índices no valor mais baixo (1,7), e índices similares para extroversão e amabilidade (1,8). O Cluster 1 possui apenas um índice comportamental mais alto que o Cluster 2, referente a uma maior demonstração pública de conteúdos que gostou (1.4x superior). Vale ressaltar que o Cluster 2 possui na comparação relativa o menor índice médio para amabilidade (2,8).

De forma sintética, agrupou-se os indicadores que foram mais significativos em cada agrupamento. Na Tabela 13, é possível perceber que o Cluster 1 possui um engajamento significativo na rede social, embora menor em relação ao Cluster 2, que destaque-se por ser um grupo mais extrovertido, amável e consciente, com índices menores de neuroticismo e amabilidade quando comparado contra si, ainda que maiores quando comparados ao Cluster 1 que realizou maiores demonstrações de apreciação na média.

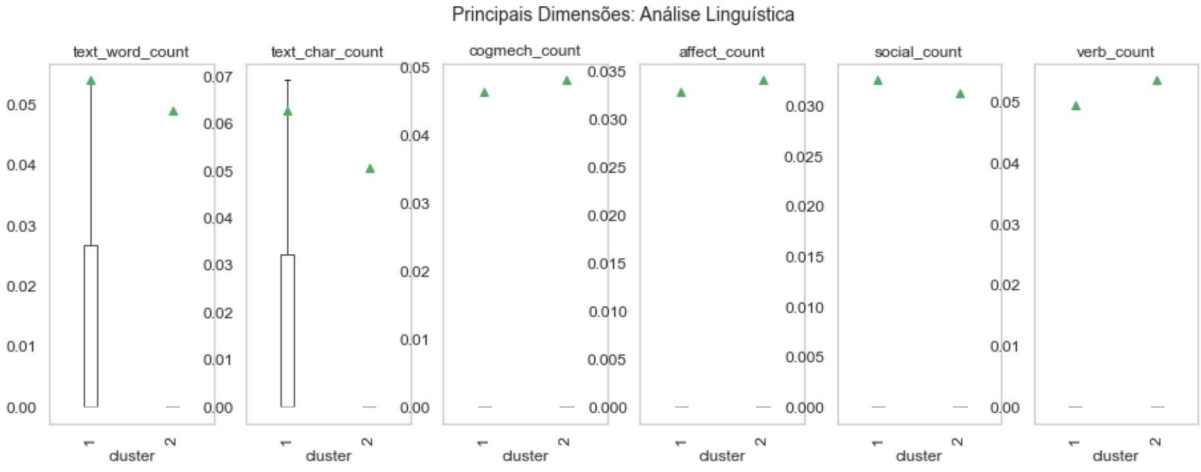

Figura 49: Análise de distribuição dos clusters no conjunto Principais Dimensões: Análise Linguística

	Cluster 1	**Cluster 2**
Engajamento	Médio	Alto
Demonstrações de apreciação	Alto	Médio
Conscienciosidade	Mais Baixa	Levemente menor
Neuroticismo	Alto	Alto
Extroversão	Mais Baixa	Alta
Amabilidade	Mais Baixa	Alta
Abertura à Experiências	Mais Baixa	Mais alta

Tabela 13: Representação sintética das características de cada Cluster.

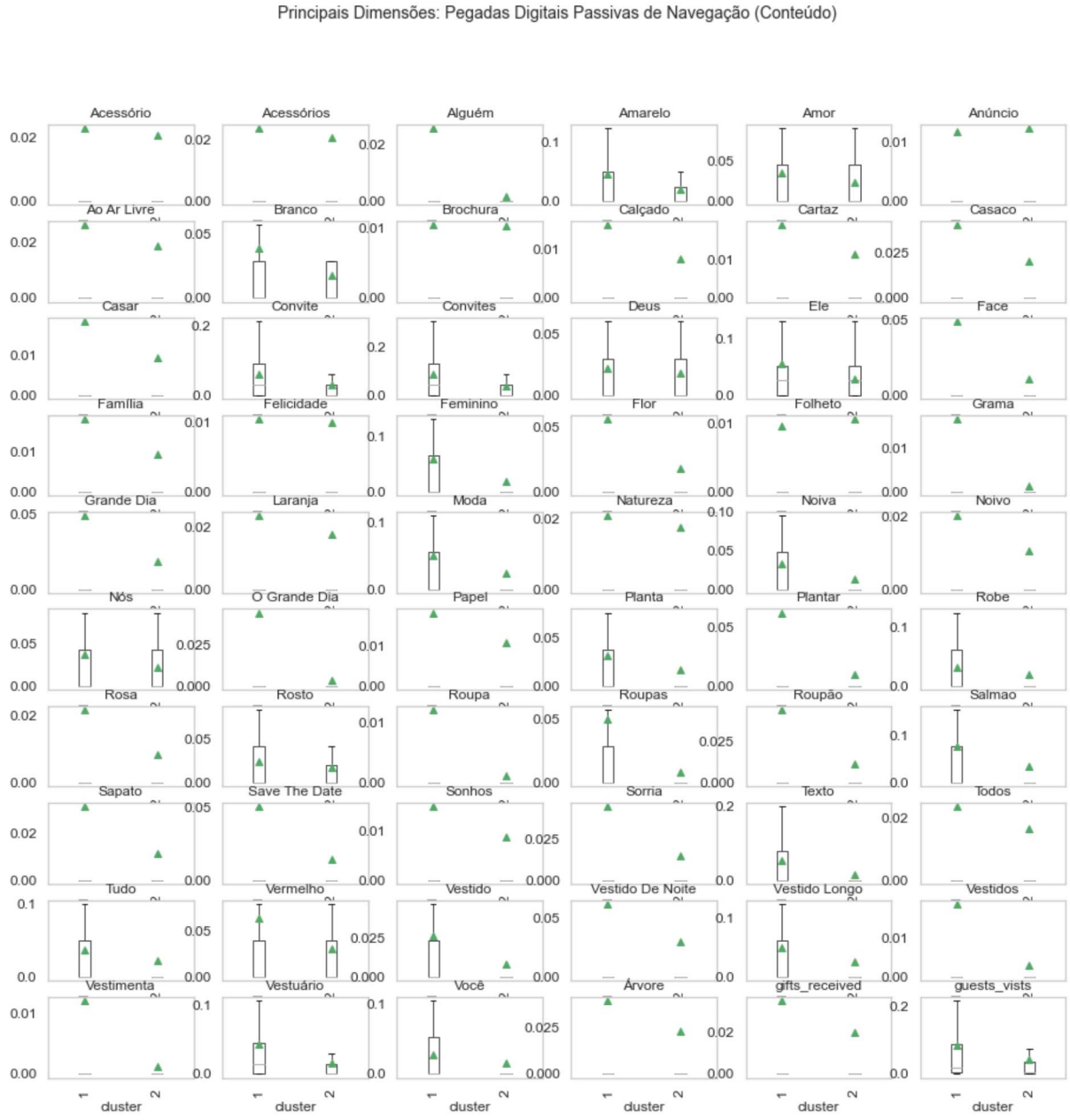

Figura 50: Análise de distribuição dos clusters no conjunto Principais Dimensões: Pegadas Digitais Passivas de Comportamento (categoria de conteúdo)

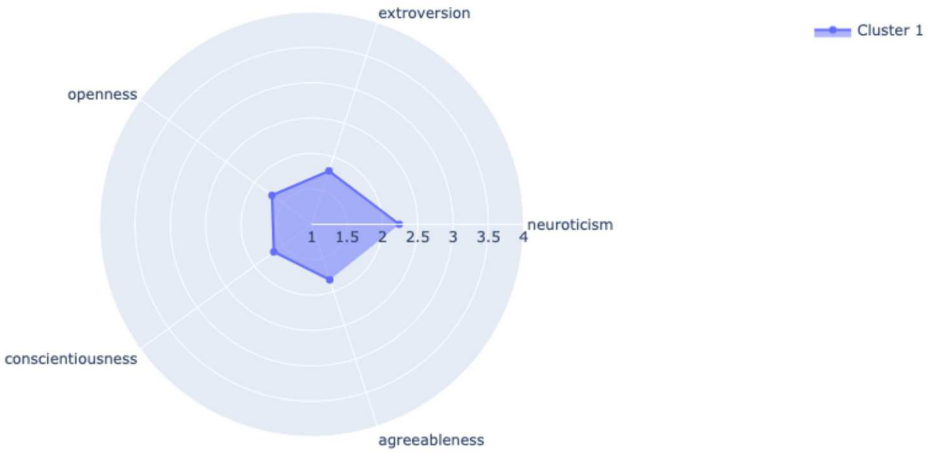

Figura 51: Distribuição dos traços de personalidade no modelo dos Cinco Grande Fatores para o Cluster 1

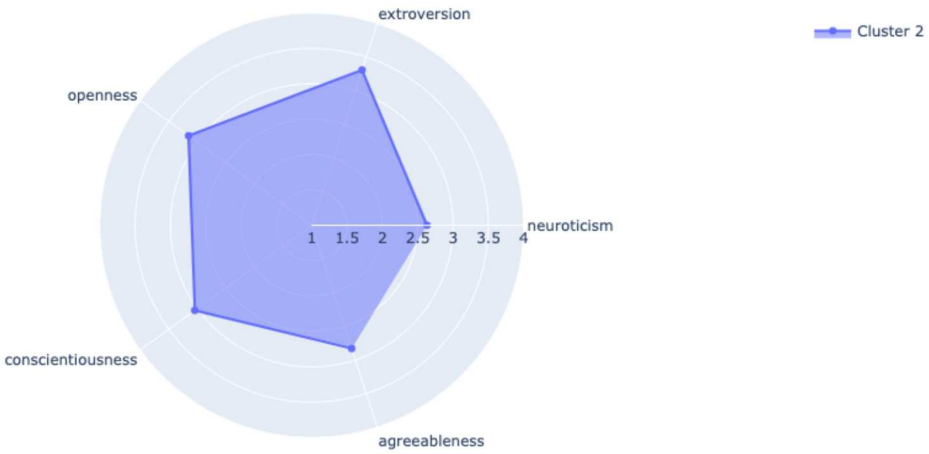

Figura 52: Distribuição dos traços de personalidade no modelo dos Cinco Grande Fatores para o Cluster 2

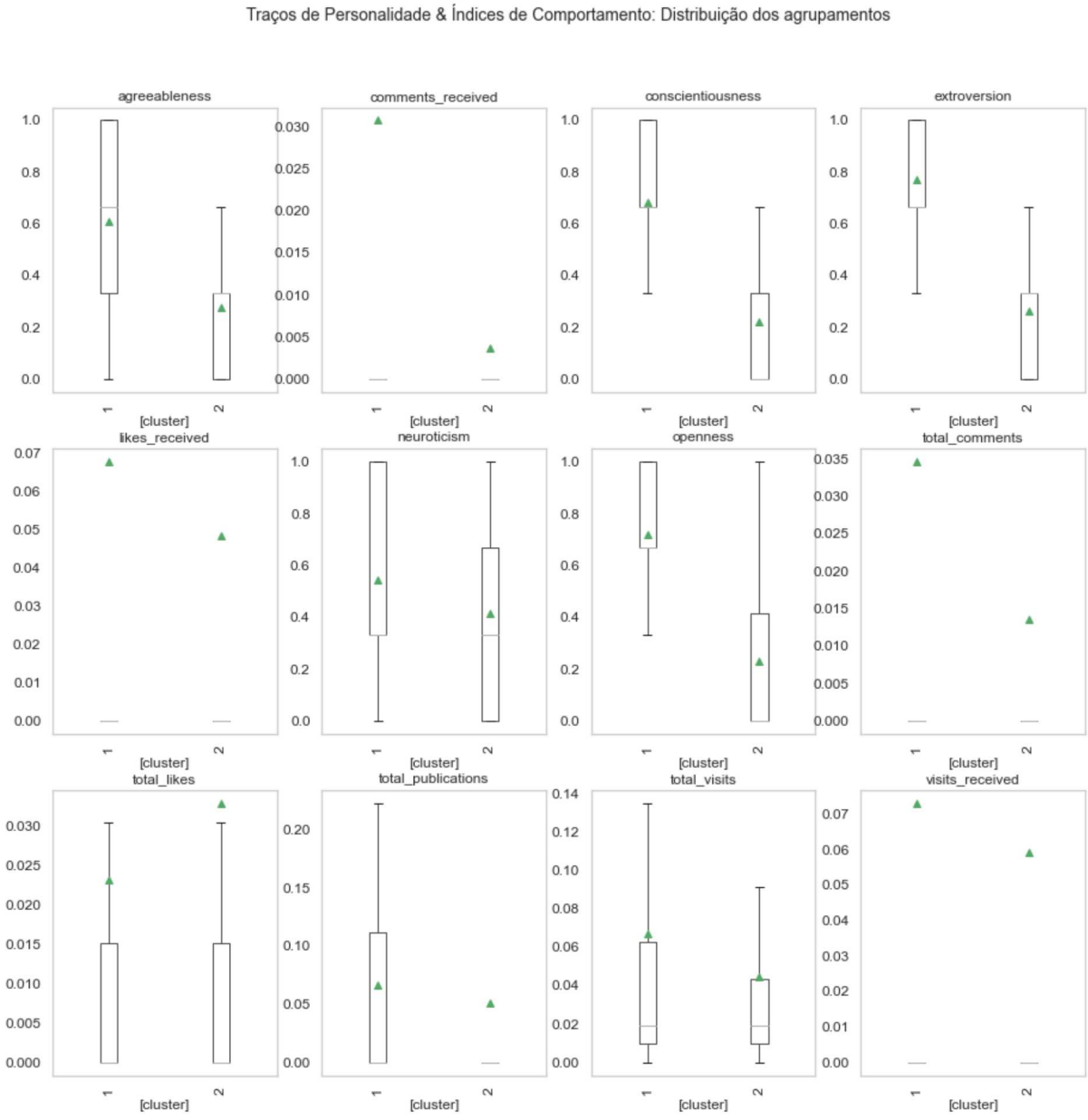

Figura 53: Traços de Personalidade Índices de Comportamento: Distribuição dos agrupamentos

7.5 Discussão

O trabalho descrito no presente documento propõe uma alternativa de estudo sobre o comportamento humano e da personalidade a partir de pegadas digitais ativas e passivas deixadas por usuários de redes sociais. Na alternativa aqui apresentada, observou-se de forma descritiva, embora em um conjunto de dados pequeno e com alta dimensionalidade, o desenvolvimento de agrupamentos utilizando métodos indiretos de aprendizado de máquina não supervisionado. Esses agrupamentos, como diferencial ótico em relação aos demais trabalhos relacionados e a partir do acesso irrestrito a conjunto de dados distintos, foram preparados, desenvolvidos e estudados com ênfase na observação de dimensões comportamentais e demográficos originadas tanto de pegadas digitais ativas quanto passivas, e da coleta permissiva de dados socioafetivos, especificamente traços de personalidade do modelo dos Cinco Grandes Fatores. Buscou-se igualmente analisar estatisticamente eventuais correlações da personalidade com os comportamentos e demais características comuns a potenciais agrupamentos distintos de usuários.

O estudo da detecção ou da observação de dados socioafetivos usando métodos indiretos é de interesse a numerosos campos, incluindo, mas não limitados a, psicologia, computação afetiva, computação da personalidade, desenvolvimento de recursos humanos e outros. Ao mesmo tempo, esse é também de interesse comercial das empresas digitais com maior valor de mercado e grandes conjuntos de dados coletados diariamente com representatividade elevada da população global e alvo de disputas de interesse social para a proteção da privacidade individual. No entanto, pode-se observar que ambas as vertentes de aplicação, seja acadêmia ou comercial, representam certas limitações e, potencialmente, ambas limitadas ao volume de dados e permissões de acessos a uma biblioteca de pegadas digitais em volumes relevantes, principalmente quando refere-se a pegadas digitais passivas e o zelo necessário ao acesso à dados tão sensíveis a nível acadêmico por parte das corporações que atuam no segmento de redes sociais no mercado global.

Este trabalho apresentou uma abordagem que explorou uma coleta de traços de personalidade rápidas, usando questionários breves de auto-relato, e um acesso responsável à dados anonimizados de produto digital com uma rede social ativa. Dessa forma, pode-se explorar e entender se dados de personalidade poderiam ser agrupados de forma valiosa quando colocados de forma igualitária à dados comportamentais e demográficos, e ambos, novamente, explorando tanto pegadas digitais ativas quanto passivas.

Para isso, percorreu-se um caminho de mineração de dados, que começou pelo estudo dos dados coletados, o pré-processamento deles, a aplicação de algoritmos de clusterização e a análise descritiva apresentada anteriormente, utilizando-se, inclusive de aprendizado de máquina supervisionado para apuração da importância das características representadas em estratégias de clusterização em matriz de alta esparsidade. Ao mesmo tempo que o trabalho foi impactado pela pandemia global da COVID-19 e restrito em sua discussão final por hipóteses mais confiáveis nos intervalos de dados analisados e potenciais relacionamentos entre as diferentes

características analisadas para os objetivos dessa pesquisa.

De toda forma, utilizando-se de alternativas estratégicas de clusterização e suas parametrizações escolhas intrínsecas, pode-se verificar algumas correlações observadas na bibliografia revisada nesse trabalho, a partir do uso de técnicas de agrupamento para generalizar os resultados analisando 157 usuários únicos. Em ambas as estratégias principais e descritas no trabalho, pode-se perceber que usuários com pontuações mais altas no traço de conscienciosidade mostraram características que representam ser mais detalhistas no seu planejamento do casamento, ao contrário daquelas com um índice inferior nesse traço, que numa análise descritiva pareciam organizar seus eventos de forma mais improvisada. Isso está diretamente relacionado as características que definem o fator de conscienciosidade, revisado na Seção 2.3, que indica a preferência por uma abordagem organizada da vida, em contraste com uma abordagem espontânea e também nos achados decorridos em uma das questões principais designada para entender "Qual o impacto de traços de personalidade no comportamento de um usuário de redes sociais?"(Seção 5.1.1), onde estudos relacionados corroboram que usuários de redes sociais com índices altos de conscienciosidade são mais meticulosas nas suas ações. É o caso do estudo de McCrae e Costa (2003) que reforça que o comportamento de pessoas conscienciosas são movidas pela racionalidade e pela busca da excelência, onde normalmente planejam com antecedência e [pensam] com cuidado antes de agir. Ao mesmo tempo, usuários presentes em agrupamentos com comportamento de maior engajamento e retenção nas redes sociais possuíam índices mais altos de extroversão, o que também está relacionado ao fator na descrição do Modelo dos Cinco Grandes Fatores, onde a extroversão está relacionada a busca de estímulos no mundo externo, a companhia de outros e a expressão de emoções positivas, amigáveis e socialmente ativas, ou seja, pessoas que não se importam de estar no centro das atenções, e concomitante são mais suscetíveis a clicaram em anúncios de produtos e realizar mais compras quando os conteúdos são adaptados a esse traço de personalidade, como revisto em outra questão principal definida como "A detecção da personalidade como ferramenta para influenciar o comportamento de usuários nas redes sociais", e documentado por Kalish e Robins (2006) , onde indivíduos extrovertidos são mais propensos a criarem oportunidades para interações.Por outro lado, usuários com traços inferiores de abertura à experiência apresentaram índices mais baixos de exploração de conteúdos, seguindo a tendência descrita pela psicologia de serem mais convencionais e tradicionais em suas perspectivas e comportamento (MCCRAE; COSTA, 2003). De forma global, foi percebido que o traço de neuroticismo distribuído de forma balanceada pelas estratégias de agrupamento utilizadas, sendo presente de forma similar entre os agrupamentos gerados e limitando qualquer correlação relevante desse traço com os agrupamentos gerados. Um ponto a ser observado no estudo, é que a amabilidade, como revisto na literatura aqui apresentada, e documentado na pesquisa de Freitag e Bauer (2016), refere-se a capacidade de manter relações sociais positivas, ser amigável, compassivo e cooperativo e que está relacionada diretamente ao altruísmo, a bondade, ao afeto e outros comportamentos pró-sociais, não apresentou índices mais altos no agrupamento que demonstrou mais atos de

apreciação públicas aos demais usuários. Embora o fator de demonstração de apreciação não pode ser considerado baixo, e no caso classificado como médio, quando comparado ao grupo que realizou mais ações de demonstração, ele é relativamente inferior, ao mesmo tempo que está diretamente correlacionado ao grupo que possui maior índice de extroversão que pode ter um comportamento esperado similar ao pressuposto aqui de realizar mais interações sociais.

Na exploração de técnicas de agrupamentos, observou-se também que o algoritmo escolhido primariamente, no caso o K-Means, não teve uma performance soberana nos índices analisando quando comparado a outros algoritmos clássicos baseados em agrupamentos pré-definidos. Dessa forma essa pesquisa também aprofundou-se no estudo de outros algoritmos, destacando-se os resultados apresentados também pelo algoritmo Spectral Clustering. Ressalta-se também que dentro das quatro estratégias de algoritmos distintos, os agrupamentos eram formados por apenas dois clusters em 50% dos casos, e em 75% dos casos considerando as estratégias com 3 clusters, onde um deles possuía apenas uma amostra.

Como referido anteriormente, devido ao pequeno conjunto de dados observados e ao curto comportamento apresentado pelos usuários da rede sociais impactados pelas questões sanitárias e o estímulo delas na utilização da rede social observada, pode-se notar que as implicações atuais tem pouca relevância prática e acadêmica. Porém, a metodologia aplicada, mostra iniciativas relevantes para serem observadas em estudos de maior escala, como o questionário de breve-relato que teve aceitação do público-alvo do estudo - ao contrário dos trabalhos relacionados que utilizavam-se de questionários longos de inferência de personalidade, a coleta de pegadas digitais passivas com alta dimensionalidade de informações relevantes, e a metodologia utilizada de mineração de dados para estratégias de clusterização aplicada em matrizes esparsas. Os resultados do estudo também sugerem uma série de outras linhas promissoras de pesquisa sobre o uso de aprendizagem de máquina em novas estratégias de clusterização, predição de personalidade via pegadas digitais passivas e o estudo de sistemas de recomendação baseado em segmentação de usuários de redes sociais. Além da potencial utilização desse conjunto de dados, com mais lastro de coleta, para estudos incrementais dos objetivos dispostos pela pesquisa aqui apresentada.

Ainda assim, este trabalho apresenta o primeiro a explorar agrupamentos com tantas dimensões de origens distintas com enfase em pegadas digitais passivas e utilizando de questionários curtos de auto-relato em uma rede social de língua portuguesa. Conforme apresentado na Tabela 1, que apresenta uma comparação dos trabalhos relacionados, a exploração aqui apresentada não conhece um método padrão comparativo para análise da eficácia das técnicas utilizadas para o desenvolvimento dos agrupamentos, onde os demais trabalhos estão relacionados a detecção da personalidade a partir de dados de treinamento rotulados.

8 CONCLUSÕES

Este trabalho apresentou os desafios do estudo de Computação da Personalidade em contextos de redes sociais comerciais de grande escala, com importantes restrições de acesso à dados sensíveis. Embora muitas são as pegadas digitais ativas e passivas deixadas por usuários de redes sociais, o interesse acadêmico em avançar em campos relacionados ao comportamento humano e assuntos como o reconhecimento de personalidade automática, percepção e síntese, esbarra em acessar irrestritamente a essas informações geradas massivamente. Isso é ainda mais relevante, quanto se trata das pegadas digitais passivas, coletadas e armazenadas pelas proprietárias das redes sociais, sem consentimento explícito aos seus usuários.

Nesse contexto, a pesquisa aqui descrita dedicou-se a criar e analisar agrupamentos de usuários de redes pelo conjunto de suas pegadas digitais ativas e passivas de suas atividades em redes sociais (comportamento) e características demográficas em conjunto com atributos socioafetivos (traços de personalidade), esses coletados de modo direto a partir de questionários curtos de auto-relato no modelo dos Cinco Grande Fatores. Com isso, derivou-se dois objetivos principais onde (1) o desenvolvimento de agrupamentos, a partir de técnicas de Mineração de Dados, considerando comportamento, personalidade e dados demográficos, permitiu a verificação da possibilidade de criação de grupos significativos considerando características socioafetivas e pegadas digitais passivas, e a consequente (2) análise qualitativa e quantitativa dos grupos produzidos, a fim de entender a qualidade dos grupos formados e a validade deles em relação aos conhecimentos revisados da Psicologia da Personalidade.

Em relação ao primeiro objetivo, verificou-se, a viabilidade da formação de grupos significativos utilizando uma metodologia que colocou todas as dimensões, de uma matriz esparsa, em um volume raso de dados, lado a lado com os grupos formados com pontuações verificadas a partir de métricas de análise de qualidade, como a verificação da distância e dispersão dos grupos formados, em uma análise que comparou 96 estratégias de agrupamentos distintos. Ressalta-se que, na metodologia aplicada, o algoritmo escolhido preferencialmente, no caso o K-Means, não teve uma performance soberana nos índices analisados quando comparado a outros algoritmos clássicos particionados e hierárquicos, destacando-se o Spectral Clustering. Observou-se também que dentro das quatro estratégias de algoritmos distintos, os agrupamentos eram formados por apenas dois clusters em 50% dos casos, e em 75% dos casos considerando as estratégias com 3 clusters, com desafios evidentes de trabalhar com dados considerados outliers.

O segundo objetivo, que dedicou-se a analisar a conexão e a segmentação de informações sensíveis de personalidade, de comportamento e de demografia em um conjunto único de dados a fim de explorar padrões a partir de agrupamentos gerados por diferentes algoritmos e estratégias de seleção de características, encontrou indícios observacionais sobre algumas características comportamentais observadas na bibliografia revisada nesse trabalho sobre Personalidade. Esse foi o caso de usuários com índices mais altos de conscienciosidade que mostraram características que representam mais criteriosidade no planejamento de eventos, ao contrário daqueles

com valores inferiores que planejavam os eventos de forma improvisada. Ao mesmo tempo, usuários com índices inferiores no traço de abertura à experiência exploraram menos novos conteúdos em relação aqueles mais propensos a novas descobertas.

Embora esse trabalho, obviamente, careça de apoio experimental, ele é um primeiro passo para propostas futuras a fim de trazer consciência sobre a relação das redes sociais, a Computação da Personalidade e os diversos campos subjacentes acadêmicos e comerciais relacionados a dados estritamente pessoais e sensíveis. Como a área da Computação da Personalidade está em constante crescimento acadêmico e interesse comercial, embora seja ainda uma área recente e que com amplitude de descobertas, pesquisas exploratórias como essa, mesmo em menor escala, podem apresentar direcionamentos e sugerir novas diretrizes, como definições de metodologias comparativas e uma possível fonte de dados para análises descritivas comparativas com base em mineração de dados.

Como próximos passos, essa pesquisa deve prosseguir e aprofundar-se em estudos exploratórios não documentados sobre a análise estatística correlacional da personalidade com os comportamentos originados de pegadas digitais ativas e passivas. Com essa abordagem, é plausível gerar novas contribuições no entendimento de como a personalidade pode potencialmente ser sintetizada por padrões comportamentais e informações demográficas processados a partir da coleta de pegadas digitais ativas. Além disso, com uma expectativa de retomada de aumento de usuários ativos na rede social e na sua densidade de uso, a coleta de dados permanecerá ativa em um fluxo contínuo para a validação dos achados preliminares desse trabalho.

8.1 Ameaças à Validade dos Resultados

Enquanto a maioria dos trabalhos relacionados é baseada em amostras limitadas (e geralmente homogêneas) e leva a alguns resultados contraditórios, potencialmente esse trabalho buscou explorar um conjunto de dados acessível e viável a ponto de ser uma alternativa para novas pesquisas serem desenvolvidas na área.

Isso ocorre porque os dados rotulados com tipos de personalidade geralmente são caros e demorados para serem coletados. A abordagem para lidar com essa limitação é entender qual é o questionário mínimo aplicável com validade estatística para criar o maior conjunto de dados possível. O questionário proposto pode ser visto no Apêndice A.1.

Os dados coletados podem sofrer um viés de domínio. Isso significa que a rede social, por estar conectada diretamente ao contexto de planejamento de casamentos, possui algumas características particulares, como a ausência de conexões diretas entre os usuários, os seus conteúdos gerados são restritivos ao contexto de casamentos e a frequência de uso e engajamento são inferiores àquelas encontradas nas redes sociais mais populares. Além disso, o público analisado é bem restrito: apenas pessoas que estão organizando seus casamentos no Brasil, com acesso à Internet e utilizando um Sistema de Informação para auxiliar no processo.

Outra limitação razoável é a quantidade de dados analisadas. Embora a quantidade de atri-

butos seja grande, os dados são relativamente esparsos. Isto limita consideravelmente a análise. Outra coisa que pode ser importante é delimitar melhor as características do público desta rede social. Essa limitação do estudo aparece de forma contemporânea ao surgimento da pandemia causada pela COVID-19 (VELAVAN; MEYER, 2020) que afetou diretamente a coleta de dados devido exclusivamente ao domínio da rede social estudada que refere-se a eventos de casamentos e as restrições de distanciamento social impostas pela pandemia.

8.2 Trabalhos Futuros

Como trabalhos futuros, novas oportunidades podem ser exploradas a medida que esse conjunto de dados inicial tenha mais volumes de dados em relação a usuários com traços de personalidade inferido e pegadas digitais com profundidade de atividades realizadas na rede social. Porém, é primário, que essa pesquisa seja revisitada seguindo a metodologia aqui proposta, quando esse conjunto de dados atingir um maior grau de maturidade e o trabalho com uma matriz tão esparsa de dados se torne estatisticamente significativa.

Dado esse volume potencial de coleta de dados futuro, como próximos passos, essa pesquisa deve prosseguir e aprofundar-se em estudos exploratórios não documentados sobre a análise estatística correlacional da personalidade com os comportamentos originados de pegadas digitais ativas e passivas. Com essa abordagem, é plausível gerar novas contribuições no entendimento de como a personalidade pode potencialmente ser sintetizada por padrões comportamentais e informações demográficas processados a partir da coleta de pegadas digitais ativas. Além disso, com uma expectativa de retomada de aumento de usuários ativos na rede social e na sua densidade de uso, a coleta de dados permanecerá ativa em um fluxo contínuo para a validação dos achados preliminares desse trabalho.

Com o maior volume de dados e uma navegação com maior profundidade dos usuários da rede social, a pesquisa pode avançar na realização de análises mais valiosas com perfis de usuários mais robustos, estendendo, inclusive, a gama de dimensões usadas neste trabalho. Além disso, análises estatísticas correlacionais da personalidade com os comportamentos é o campo a ser explorado, a fim de entender as dependências entre as características socio-afetivas de cada usuário e o reflexo nos seus comportamentos online.

À vista disso, entre potenciais novas oportunidades consequentes do posterior avanço dos estudos aqui documentados, pode-se destacar três principais caminhos a percorrer:

1. **Sistemas de recomendação:** O massivo interesse em redes sociais e a personalidade social intrinsecamente relacionada nas pessoas, nas relações entre elas e nos seus comportamentos e hábitos dentro desses ambientes digitais habilita uma série de tópicos relevantes para a otimização de sistemas de recomendação. A Computação de Personalidade permite entender as preferências dos usuários e criar sistemas de recomendação de uma perspectiva diferente. Desa forma, aproveitar os traços de personalidades dos usuários para melhorar as recomendações pode ser eficaz na solução de problemas clássicos de

sistemas convencionais de recomendação. Tais como o problema de inicialização a frio (*cold-start problem*), quando o sistema não tem muitos dados sobre as preferências do usuário e problemas relacionados a esparsidade de dados. No entanto, com a compreensão da personalidade do usuário, há de se ressaltar o desafio e oportunidade de compreender ainda mais o poder dos casos de uso dos sistemas de recomendação dentro redes sociais para influenciar o comportamento de usuários nas redes sociais para fins políticos e comerciais, como revisto na Seção 5.1.6.

2. **Novas estratégias de clusterização:** A partir da natural evolução dessa pesquisa, com um volume maior de dados coletados, novas estratégias e técnicas de clusterização podem ser aplicadas, onde o método proposto e desenvolvido nessa pesquisa pode ser alterado em algumas facetas. Uma das possibilidades de exploração de clusterização é a extração de características a partir das mídias ricas publicadas na rede social, como fotos e vídeos, e a relação dos usuários nas redes sociais a partir de suas interações (grafos de relacionamento). Outros algoritmos também podem ser aplicados, com um maior volume de dados, como aqueles baseados em Modelos de Mistura Gaussiana (*GMMs*), pois apresenta um nível mais alto de flexibilidade em relação à covariância de cluster em comparação com o cluster K-means. Como esse conceito usa probabilidade, é possível analisar do ponto de vista probabilístico cada amostra de usuário, possibilitando a distribuição proporcional de um usuário em mais de um cluster. A Análise de Componentes Principais (*PCA*) para redução de dimensionalidade é uma abordagem a ser considerada para uma visualização mais rica dos agrupamentos gerados, facilitando a interpretação dos resultados e potenciais ajustes-finos ao reduzir a complexidade do modelo. O desempenho dos modelos desenvolvimentos também pode ser acrescido de novos índices como Dunn, PBM e outros, como critério de comparação.

3. **Predição de personalidade via pegadas digitais passivas:** O conjunto de dados anonimizado e rotulado que foi produzido e disponibilizado por essa pesquisa, pode ser utilizado para o desenvolvimento de algoritmos de aprendizado de máquina com o objetivo de detectar e avaliar traços de personalidade de usuários de redes sociais permitindo a elaboração de novas arquiteturas e processos para melhorar sistemas de detecção da personalidade, contribuindo para o estado da arte, revisado na Seção 5.1.3, ao adicionar pegadas digitais passivas, um domínio relativamente novo na pesquisa de aprendizado de máquina com ênfase no estudo da personalidade através de redes sociais.

REFERÊNCIAS

ACKERMANN, M. R.; BLÖMER, J.; KUNTZE, D.; SOHLER, C. Analysis of agglomerative clustering. **Algorithmica**, [S.l.], v. 69, n. 1, p. 184–215, 2014.

ADALI, S.; GOLBECK, J. Predicting personality with social behavior. In: IEEE/ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING, 2012., 2012. **Anais...** [S.l.: s.n.], 2012. p. 302–309.

ADLAI-GAIL, W. S. Exploring the autotelic personality. , [S.l.], 1995.

AGGARWAL, C. C. **Data mining**: the textbook. [S.l.]: Springer, 2015.

ALAM, F.; STEPANOV, E. A.; RICCARDI, G. Personality traits recognition on social network-facebook. In: SEVENTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2013. **Anais...** [S.l.: s.n.], 2013.

ANDRADE, J. M. de. **Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil**. 2008. Tese (Doutorado em Ciência da Computação) — Ph. D. thesis, Universidade de Brasília, 2008.

APPLING, D. S.; BRISCOE, E. J.; HAYES, H.; MAPPUS, R. L. Towards automated personality identification using speech acts. In: SEVENTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2013. **Anais...** [S.l.: s.n.], 2013.

ARAKERIMATH, A. R.; GUPTA, P. K. Digital Footprint: pros, cons, and future. **International Journal of Latest Technology in Engineering**, [S.l.], v. 4, n. 10, p. 52–56, 2015.

ARNOLD, M. B. Emotion and personality. **Columbia University Press**, [S.l.], v. 1, 1960.

ARORA, P.; VARSHNEY, S. et al. Analysis of k-means and k-medoids algorithm for big data. **Procedia Computer Science**, [S.l.], v. 78, p. 507–512, 2016.

AZUCAR, D.; MARENGO, D.; SETTANNI, M. Predicting the Big 5 personality traits from digital footprints on social media: a meta-analysis. **Personality and Individual Differences**, [S.l.], v. 124, p. 150–159, 2018.

BACHRACH, Y.; KOSINSKI, M.; GRAEPEL, T.; KOHLI, P.; STILLWELL, D. Personality and patterns of Facebook usage. In: ACM WEB SCIENCE CONFERENCE, 4., 2012. **Proceedings...** [S.l.: s.n.], 2012. p. 24–32.

BACK, M. D.; STOPFER, J. M.; VAZIRE, S.; GADDIS, S.; SCHMUKLE, S. C.; EGLOFF, B.; GOSLING, S. D. Facebook profiles reflect actual personality, not self-idealization. **Psychological science**, [S.l.], v. 21, n. 3, p. 372–374, 2010.

BAI, S.; GAO, R.; ZHU, T. Determining personality traits from renren status usage behavior. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL VISUAL MEDIA, 2012. **Anais...** [S.l.: s.n.], 2012. p. 226–233.

BARBIER, G.; LIU, H. Data mining in social media. In: **Social network data analytics**. [S.l.]: Springer, 2011. p. 327–352.

BERKHIN, P. A survey of clustering data mining techniques. In: **Grouping multidimensional data**. [S.l.]: Springer, 2006. p. 25–71.

BHAVYA, S.; PILLAI, A. S.; GUAZZARONI, G. Personality Identification from Social Media Using Deep Learning: a review. In: **Soft Computing for Problem Solving**. [S.l.]: Springer, 2020. p. 523–534.

BOUTSIDIS, C.; ZOUZIAS, A.; MAHONEY, M. W.; DRINEAS, P. Randomized dimensionality reduction for k-means clustering. **IEEE Transactions on Information Theory**, [S.l.], v. 61, n. 2, p. 1045–1062, 2014.

BURBACH, L.; HALBACH, P.; ZIEFLE, M.; CALERO VALDEZ, A. Who Shares Fake News in Online Social Networks? In: ACM CONFERENCE ON USER MODELING, ADAPTATION AND PERSONALIZATION, 27., 2019. **Proceedings...** [S.l.: s.n.], 2019. p. 234–242.

CELEBI, M. E.; KINGRAVI, H. A.; VELA, P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. **Expert systems with applications**, [S.l.], v. 40, n. 1, p. 200–210, 2013.

CHAPSKY, D. Leveraging online social networks and external data sources to predict personality. In: INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING, 2011., 2011. **Anais...** [S.l.: s.n.], 2011. p. 428–433.

CHARLESWORTH, A. **An introduction to social media marketing**. [S.l.]: Routledge, 2014.

CHAUHAN, D.; MATHEWS, R. Review on Dimensionality Reduction Techniques. In: INTERNATIONAL CONFERENCE ON COMPUTER NETWORKS, BIG DATA AND IOT, 2019. **Anais...** [S.l.: s.n.], 2019. p. 356–362.

CHEN, Y.-J.; CHEN, Y.-M.; HSU, Y.-J.; WU, J.-H. Predicting Consumers' Decision-Making Styles by Analyzing Digital Footprints on Facebook. **International Journal of Information Technology & Decision Making**, [S.l.], v. 18, n. 02, p. 601–627, 2019.

CHEN, Y.; PAVLOV, D.; CANNY, J. F. Large-scale behavioral targeting. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 15., 2009. **Proceedings...** [S.l.: s.n.], 2009. p. 209–218.

CHIN, D. N.; WRIGHT, W. R. Social Media Sources for Personality Profiling. In: UMAP WORKSHOPS, 2014. **Anais...** [S.l.: s.n.], 2014.

COHEN, M. B.; ELDER, S.; MUSCO, C.; MUSCO, C.; PERSU, M. Dimensionality reduction for k-means clustering and low rank approximation. In: ACM SYMPOSIUM ON THEORY OF COMPUTING, 2015. **Proceedings...** [S.l.: s.n.], 2015. p. 163–172.

COLTHEART, M. The MRC psycholinguistic database. **The Quarterly Journal of Experimental Psychology Section A**, [S.l.], v. 33, n. 4, p. 497–505, 1981.

DAMASIO, A. R. Descartes' error: emotion, rationality and the human brain. **New York: Putnam**, [S.l.], p. 1061–1070, 1994.

DASGUPTA, S. Experiments with random projection. **arXiv preprint arXiv:1301.3849**, [S.l.], 2013.

DEYOUNG, C. G.; GRAY, J. R. Personality neuroscience: explaining individual differences in affect, behavior, and cognition. **The Cambridge Handbook of Personality Psychology**, [S.l.], p. 323–346, 2009.

DUTT, A.; ISMAIL, M. A.; HERAWAN, T. A systematic review on educational data mining. **Ieee Access**, [S.l.], v. 5, p. 15991–16005, 2017.

DWORK, C.; MULLIGAN, D. K. It's not privacy, and it's not fair. **Stan. L. Rev. Online**, [S.l.], v. 66, p. 35, 2013.

EXAME. **Startup brasileira que monta casamentos já quer internacionalizar**. (Accessed on 05/17/2020), `https://exame.abril.com.br/pme/esta-startup-planeja-seu-casamento-e-ja-vai-internacionalizar/`.

FARNADI, G.; ZOGHBI, S.; MOENS, M.-F.; DE COCK, M. Recognising personality traits using Facebook status updates. In: SEVENTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2013. **Anais...** [S.l.: s.n.], 2013.

FILIPPOVA, K. User demographics and language in an implicit social network. In: JOINT CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL NATURAL LANGUAGE LEARNING, 2012., 2012. **Proceedings...** [S.l.: s.n.], 2012. p. 1478–1488.

FISH, T. **My Digital Footprint A two-sided digital business model where your privacy will be someone else's business!** [S.l.]: futuretext, 2009.

FREITAG, M.; BAUER, P. C. Personality traits and the propensity to trust friends and strangers. **The social science journal**, [S.l.], v. 53, n. 4, p. 467–476, 2016.

FREITAS, A. R. R.; NAPIMOGA, M.; DONALISIO, M. R. Análise da gravidade da pandemia de Covid-19. **Epidemiologia e Serviços de Saúde**, [S.l.], v. 29, p. e2020119, 2020.

GARCÍA, S.; RAMÍREZ-GALLEGO, S.; LUENGO, J.; BENÍTEZ, J. M.; HERRERA, F. Big data preprocessing: methods and prospects. **Big Data Analytics**, [S.l.], v. 1, n. 1, p. 1–22, 2016.

GAVRILOVA, M. L. Machine Learning for Social Behavior Understanding. In: **Proceedings of Computer Graphics International 2018**. [S.l.: s.n.], 2018. p. 247–252.

GOLBECK, J.; ROBLES, C.; EDMONDSON, M.; TURNER, K. Predicting personality from twitter. In: IEEE THIRD INTERNATIONAL CONFERENCE ON PRIVACY, SECURITY, RISK AND TRUST AND 2011 IEEE THIRD INTERNATIONAL CONFERENCE ON SOCIAL COMPUTING, 2011., 2011. **Anais...** [S.l.: s.n.], 2011. p. 149–156.

GOLBECK, J.; ROBLES, C.; TURNER, K. Predicting personality with social media. In: **CHI'11 extended abstracts on human factors in computing systems**. [S.l.: s.n.], 2011. p. 253–262.

GOLDER, S. A.; MACY, M. W. Digital footprints: opportunities and challenges for online social research. **Annual Review of Sociology**, [S.l.], v. 40, p. 129–152, 2014.

GÖSSLING, S.; SCOTT, D.; HALL, C. M. Pandemics, tourism and global change: a rapid assessment of covid-19. **Journal of Sustainable Tourism**, [S.l.], v. 29, n. 1, p. 1–20, 2020.

GRAHAM, J. R. **MMPI-2**: assessing personality and psychopathology. [S.l.]: Oxford University Press, 1990.

GRANVILLE, K. Facebook and Cambridge Analytica: what you need to know as fallout widens. **The New York Times**, [S.l.], v. 19, 2018.

HALIM, Z.; ATIF, M.; RASHID, A.; EDWIN, C. A. Profiling players using real-world datasets: clustering the data and correlating the results with the big-five personality traits. **IEEE Transactions on Affective Computing**, [S.l.], 2017.

HAMERLY, G.; ELKAN, C. Learning the k in k-means. **Advances in neural information processing systems**, [S.l.], v. 16, p. 281–288, 2004.

HANCOCK, J. T.; KHOSHGOFTAAR, T. M. Survey on categorical data for neural networks. **Journal of Big Data**, [S.l.], v. 7, p. 1–41, 2020.

HAND, D. J. Data Mining. **Encyclopedia of Environmetrics**, [S.l.], v. 2, 2006.

HAND, D. J. Principles of data mining. **Drug safety**, [S.l.], v. 30, n. 7, p. 621–622, 2007.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: a k-means clustering algorithm. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, [S.l.], v. 28, n. 1, p. 100–108, 1979.

HINDS, J.; JOINSON, A. Human and Computer Personality Prediction From Digital Footprints. **Current Directions in Psychological Science**, [S.l.], p. 0963721419827849, 2019.

HOFSTEDE, G.; HOFSTEDE, G. J.; MINKOV, M. **Cultures and organizations**: software of the mind. [S.l.]: Citeseer, 1991. v. 2.

ISAAK, J.; HANNA, M. J. User data privacy: facebook, cambridge analytica, and privacy protection. **Computer**, [S.l.], v. 51, n. 8, p. 56–59, 2018.

ISMAILI, O. A.; LEMAIRE, V.; CORNUÉJOLS, A. A supervised methodology to measure the variables contribution to a clustering. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING, 2014. **Anais...** [S.l.: s.n.], 2014. p. 159–166.

JACHIMOWICZ, J.; MATZ, S.; POLONSKI, V. The Behavioral Scientist's Ethics Checklist. **Behavioral Scientist**, [S.l.], 2017.

JAQUES, P. A. Using an animated pedagogical agent to interact affectively with the student. , [S.l.], 2004.

JAWLIK, A. A. **Statistics from a to z**: confusing concepts clarified. [S.l.]: John Wiley & Sons, 2016.

JIANG, F.; LEUNG, C. K.; PAZDOR, A. G. Big data mining of social networks for friend recommendation. In: IEEE/ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING (ASONAM), 2016., 2016. **Anais...** [S.l.: s.n.], 2016. p. 921–922.

JOHN, O. P.; DONAHUE, E. M.; KENTLE, R. L. **The big five inventory—versions 4a and 54**. [S.l.]: Berkeley, CA: University of California, Berkeley, Institute of Personality ..., 1991.

KALISH, Y.; ROBINS, G. Psychological predispositions and network structure: the relationship between individual predispositions, structural holes and network closure. **Social networks**, [S.l.], v. 28, n. 1, p. 56–84, 2006.

KAUSHAL, V.; PATWARDHAN, M. Emerging trends in personality identification using online social networks—a literature survey. **ACM Transactions on Knowledge Discovery from Data (TKDD)**, [S.l.], v. 12, n. 2, p. 1–30, 2018.

KENNEDY, H.; ELGESEM, D.; MIGUEL, C. On fairness: user perspectives on social media data mining. **Convergence**, [S.l.], v. 23, n. 3, p. 270–288, 2017.

KHAWAJA, S. G.; KHAN, A. M.; AKRAM, M. U.; KHAN, S. A. A Novel Architecture for k-means Clustering Algorithm. In: INTERNATIONAL AFRO-EUROPEAN CONFERENCE FOR INDUSTRIAL ADVANCEMENT, 2016. **Anais...** [S.l.: s.n.], 2016. p. 311–320.

KOSINSKI. **myPersonality.org - Publications**. (Accessed on 05/17/2020), `https://sites.google.com/michalkosinski.com/mypersonality/publications?authuser=0`.

KOSINSKI. **myPersonality.org**. (Accessed on 06/13/2020), `https://sites.google.com/michalkosinski.com/mypersonality`.

KOSINSKI, M.; BACHRACH, Y.; KOHLI, P.; STILLWELL, D.; GRAEPEL, T. Manifestations of user personality in website choice and behaviour on online social networks. **Machine learning**, [S.l.], v. 95, n. 3, p. 357–380, 2014.

KOSINSKI M., M. S. M. G.; HANCOCK, J. Facebook as a research tool for the social sciences. **American Psychologist**, [S.l.], 2005.

KOSINSKI, M.; MATZ, S. C.; GOSLING, S. D.; POPOV, V.; STILLWELL, D. Facebook as a research tool for the social sciences: opportunities, challenges, ethical considerations, and practical guidelines. **American Psychologist**, [S.l.], v. 70, n. 6, p. 543, 2015.

KOSINSKI, M.; MATZ, S.; GOSLING, S.; POPOV, V.; STILLWELL, D. Facebook as a Research Tool for the Social Sciences. **The American psychologist**, [S.l.], v. 70, p. 543–556, 09 2015.

KOSINSKI, M.; STILLWELL, D.; GRAEPEL, T. Private traits and attributes are predictable from digital records of human behavior. **Proceedings of the national academy of sciences**, [S.l.], v. 110, n. 15, p. 5802–5805, 2013.

KUSS, D. J.; GRIFFITHS, M. D. Online social networking and addiction—a review of the psychological literature. **International journal of environmental research and public health**, [S.l.], v. 8, n. 9, p. 3528–3552, 2011.

LAMBIOTTE, R.; KOSINSKI, M. Tracking the digital footprints of personality. **Proceedings of the IEEE**, [S.l.], v. 102, n. 12, p. 1934–1939, 2014.

LAROS, J. A.; PERES, A. J. d. S.; ANDRADE, J. M. d.; PASSOS, M. F. D. Validity evidence of two short scales measuring the Big Five personality factors. **Psicologia: Reflexão e Crítica**, [S.l.], v. 31, 2018.

LI, Y.; WU, H. A clustering method based on K-means algorithm. **Physics Procedia**, [S.l.], v. 25, p. 1104–1109, 2012.

LUTZ, C.; HOFFMANN, C. P. The dark side of online participation: exploring non-, passive and negative participation. **Information, Communication & Society**, [S.l.], v. 20, n. 6, p. 876–897, 2017.

MALDONADO, V. N.; BLUM, R. O.; BORELLI, A. **LGPD**: lei geral de proteção de dados: comentada. [S.l.]: Revista dos Tribunais, 2019.

MARENGO, D.; SETTANNI, M. Mining Facebook Data for Personality Prediction: an overview. In: **Digital Phenotyping and Mobile Sensing**. [S.l.]: Springer, 2019. p. 109–124.

MARKOVIKJ, D.; GIEVSKA, S.; KOSINSKI, M.; STILLWELL, D. J. Mining facebook data for predictive personality modeling. In: SEVENTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2013. **Anais...** [S.l.: s.n.], 2013.

MATZ, S. C.; APPEL, R. E.; KOSINSKI, M. Privacy in the age of psychological targeting. **Current opinion in psychology**, [S.l.], v. 31, p. 116–121, 2020.

MATZ, S. C.; KOSINSKI, M.; NAVE, G.; STILLWELL, D. J. Psychological targeting as an effective approach to digital mass persuasion. **Proceedings of the national academy of sciences**, [S.l.], v. 114, n. 48, p. 12714–12719, 2017.

MATZ, S.; KOSINSKI, M. Using Consumers' Digital Footprints for More Persuasive Mass Communication. **NIM Marketing Intelligence Review**, [S.l.], v. 11, n. 2, p. 18–23, 2019.

MCCRAE, R. R.; COSTA, P. T. **Personality in adulthood**: a five-factor theory perspective. [S.l.]: Guilford Press, 2003.

MOHAMAD, I. B.; USMAN, D. Standardization and its effects on K-means clustering algorithm. **Research Journal of Applied Sciences, Engineering and Technology**, [S.l.], v. 6, n. 17, p. 3299–3303, 2013.

MUHAMMAD, S. S.; DEY, B. L.; WEERAKKODY, V. Analysis of factors that influence customers' willingness to leave big data digital footprints on social media: a systematic review of literature. **Information Systems Frontiers**, [S.l.], v. 20, n. 3, p. 559–576, 2018.

MÜLLNER, D. Modern hierarchical, agglomerative clustering algorithms. **arXiv preprint arXiv:1109.2378**, [S.l.], 2011.

MYERS, I. B. The Myers-Briggs Type Indicator: manual (1962). , [S.l.], 1962.

NGUYEN, T.-H. T.; DINH, D.-T.; SRIBOONCHITTA, S.; HUYNH, V.-N. A method for k-means-like clustering of categorical data. **Journal of Ambient Intelligence and Humanized Computing**, [S.l.], p. 1–11, 2019.

NIU, D.; DY, J.; JORDAN, M. I. Dimensionality reduction for spectral clustering. In: FOURTEENTH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 2011. **Proceedings...** [S.l.: s.n.], 2011. p. 552–560.

OBAR, J. A.; OELDORF-HIRSCH, A. The biggest lie on the internet: ignoring the privacy policies and terms of service policies of social networking services. **Information, Communication & Society**, [S.l.], v. 23, n. 1, p. 128–147, 2020.

OLSHANNIKOVA, E.; OLSSON, T.; HUHTAMÄKI, J.; KÄRKKÄINEN, H. Conceptualizing big social data. **Journal of Big Data**, [S.l.], v. 4, n. 1, p. 3, 2017.

ÖNDER, I.; KOERBITZ, W.; HUBMANN-HAIDVOGEL, A. Tracing tourists by their digital footprints: the case of austria. **Journal of Travel Research**, [S.l.], v. 55, n. 5, p. 566–573, 2016.

OPHIR, Y.; ASTERHAN, C. S.; SCHWARZ, B. B. The digital footprints of adolescent depression, social rejection and victimization of bullying on Facebook. **Computers in Human Behavior**, [S.l.], v. 91, p. 62–71, 2019.

ORTIGOSA, A.; CARRO, R. M.; QUIROGA, J. I. Predicting user personality by mining social interactions in Facebook. **Journal of computer and System Sciences**, [S.l.], v. 80, n. 1, p. 57–71, 2014.

PETERSEN, K.; FELDT, R.; MUJTABA, S.; MATTSSON, M. Systematic mapping studies in software engineering. In: INTERNATIONAL CONFERENCE ON EVALUATION AND ASSESSMENT IN SOFTWARE ENGINEERING (EASE) 12, 12., 2008. **Anais...** [S.l.: s.n.], 2008. p. 1–10.

PICARD, R. W. **Affective computing**. [S.l.]: MIT Press, 1997.

QUERCIA, D.; LAMBIOTTE, R.; STILLWELL, D.; KOSINSKI, M.; CROWCROFT, J. The personality of popular facebook users. In: ACM 2012 CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK, 2012. **Proceedings...** [S.l.: s.n.], 2012. p. 955–964.

SANTINI, M. Advantages & disadvantages of k-means and hierarchical clustering (unsupervised learning). **URL: http://santini. se/teaching/ml/2016/Lect_10/10c_Unsupervise dMethods. pdf (Accesed 17.04. 2019)**, [S.l.], 2016.

SCHERER, K. R. What are emotions? And how can they be measured? **Social science information**, [S.l.], v. 44, n. 4, p. 695–729, 2005.

SCHERER, K. R.; BÄNZIGER, T.; ROESCH, E. **A Blueprint for Affective Computing**: a sourcebook and manual. [S.l.]: Oxford University Press, 2010.

SEGALIN, C.; CELLI, F.; POLONIO, L.; KOSINSKI, M.; STILLWELL, D.; SEBE, N.; CRISTANI, M.; LEPRI, B. What your facebook profile picture reveals about your personality. In: ACM INTERNATIONAL CONFERENCE ON MULTIMEDIA, 25., 2017. **Proceedings...** [S.l.: s.n.], 2017. p. 460–468.

SILVA, B. B. C. da; PARABONI, I. Personality recognition from Facebook text. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 2018. **Anais...** [S.l.: s.n.], 2018. p. 107–114.

SIMON, H. A. Motivational and emotional controls of cognition. **Psychological review**, [S.l.], v. 74, n. 1, p. 29, 1967.

SMITH, A. Public attitudes toward computer algorithms. **Pew Research Center**, [S.l.], 2018.

STARTSE. **Mecasei.com recebe aporte de R$ 800 mil de investidores anjos**. (Accessed on 05/17/2020), `https://www.startse.com/noticia/investimentos/mecasei-com-recebe-aporte-de-r-800-mil-de-investidores-anjos`.

STARTUP-BRASIL. **Mecasei.com lança a "Meeka", a primeira assistente pessoal para noivos**. (Accessed on 05/17/2020), https://www.startupbrasil.org.br/2016/04/19/me-casei-ponto-com/.

STATISTA. **Advertising expense of Facebook from 2014 to 2019**.

STATISTA. **Number of social media users worldwide from 2010 to 2021 (in billions)**. 2020.

STILLWELL, D. J.; KOSINSKI, M. myPersonality project: example of successful utilization of online social networks for large-scale social research. **American Psychologist**, [S.l.], v. 59, n. 2, p. 93–104, 2004.

TERRACCIANO, A.; SUTIN, A. R.; AN, Y.; O'BRIEN, R. J.; FERRUCCI, L.; ZONDERMAN, A. B.; RESNICK, S. M. Personality and risk of Alzheimer's disease: new data and meta-analysis. **Alzheimer's & Dementia**, [S.l.], v. 10, n. 2, p. 179–186, 2014.

TIKKINEN-PIRI, C.; ROHUNEN, A.; MARKKULA, J. EU General Data Protection Regulation: changes and implications for personal data collecting companies. **Computer Law & Security Review**, [S.l.], v. 34, n. 1, p. 134–153, 2018.

TUTAYSALGIR, E.; KARAGOZ, P.; TOROSLU, I. H. Clustering based personality prediction on turkish tweets. In: IEEE/ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS AND MINING, 2019., 2019. **Proceedings...** [S.l.: s.n.], 2019. p. 825–828.

TYLOR, E. B. **Primitive culture**: researches into the development of mythology, philosophy, religion, art, and custom. [S.l.]: Murray, 1871. v. 2.

VELAVAN, T. P.; MEYER, C. G. The COVID-19 epidemic. **Tropical medicine & international health**, [S.l.], v. 25, n. 3, p. 278, 2020.

VENTIUR. **Mecasei.com apresenta-se no Venture Fórum da ABVCAP - VENTIUR**. (Accessed on 05/17/2020), https://ventiur.net/mecasei-com-apresenta-se-no-venture-forum-da-abvcap/.

VERHOEVEN, B.; DAELEMANS, W.; DE SMEDT, T. Ensemble methods for personality recognition. In: SEVENTH INTERNATIONAL AAAI CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 2013. **Anais...** [S.l.: s.n.], 2013.

VERMA, D.; MEILA, M. A comparison of spectral clustering algorithms. **University of Washington Tech Rep UWCSE030501**, [S.l.], v. 1, p. 1–18, 2003.

VINCIARELLI, A.; MOHAMMADI, G. A survey of personality computing. **IEEE Transactions on Affective Computing**, [S.l.], v. 5, n. 3, p. 273–291, 2014.

VON LUXBURG, U. A tutorial on spectral clustering. **Statistics and computing**, [S.l.], v. 17, n. 4, p. 395–416, 2007.

WALD, R.; KHOSHGOFTAAR, T.; SUMNER, C. Machine prediction of personality from Facebook profiles. In: IEEE 13TH INTERNATIONAL CONFERENCE ON INFORMATION REUSE & INTEGRATION (IRI), 2012., 2012. **Anais...** [S.l.: s.n.], 2012. p. 109–115.

WALFISH, S. A review of statistical outlier methods. **Pharmaceutical technology**, [S.l.], v. 30, n. 11, p. 82, 2006.

WANG, Y.; KOSINSKI, M. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. **Journal of personality and social psychology**, [S.l.], v. 114, n. 2, p. 246, 2018.

WEI, H.; ZHANG, F.; YUAN, N. J.; CAO, C.; FU, H.; XIE, X.; RUI, Y.; MA, W.-Y. Beyond the words: predicting user personality from heterogeneous information. In: ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, 2017. **Proceedings...** [S.l.: s.n.], 2017. p. 305–314.

WHAITE, E. O.; SHENSA, A.; SIDANI, J. E.; COLDITZ, J. B.; PRIMACK, B. A. Social media use, personality characteristics, and social isolation among young adults in the United States. **Personality and Individual Differences**, [S.l.], v. 124, p. 45–50, 2018.

WILLIAMS, L.; PENNINGTON, D. An authentic self: big data and passive digital footprints. In: INTERNATIONAL SYMPOSIUM ON HUMAN ASPECTS OF INFORMATION SECURITY & ASSURANCE (HAISA 2018), 2018. **Anais...** [S.l.: s.n.], 2018.

YE, Z.; DU, Y.; ZHAO, L. Predicting Personality Traits of Users in Social Networks. In: INTERNATIONAL CONFERENCE ON INTELLIGENT DATA ENGINEERING AND AUTOMATED LEARNING, 2017. **Anais...** [S.l.: s.n.], 2017. p. 181–191.

YOUYOU, W.; KOSINSKI, M.; STILLWELL, D. Computer-based personality judgments are more accurate than those made by humans. **Proceedings of the National Academy of Sciences**, [S.l.], v. 112, n. 4, p. 1036–1040, 2015.

ZENG, Z.; HU, Y.; ROISMAN, G. I.; WEN, Z.; FU, Y.; HUANG, T. S. Audio-visual spontaneous emotion recognition. In: **Artifical Intelligence for Human Computing**. [S.l.]: Springer, 2007. p. 72–90.

ANEXO A MATERIAIS UTILIZADOS NA PESQUISA

A.1 Questionário dos Cinco Grande Fatores

INSTRUÇÕES. A seguir encontram-se algumas caracteristicas (afirmacoes) que podem ou nao lhe dizer respeito. Por favor, escolha um dos numeros na escala abaixo que melhor expresse sua opiniao em relacao a voce mesmo e anote no espaco ao lado de cada afirmacao. Vales ressaltar que nao existem respostas certas ou erradas. Utilize a seguinte escala de resposta

- **1**: Discordo totalmente;
- **2**: Discordo;
- **3**: Nem concordo nem discordo;
- **4**: Concordo;
- **5**: Concordo totalmente.

Eu me vejo como alguem que...

1. É conversador, comunicativo.
2. É minucioso, detalhista no trabalho, no que faz.
3. Insiste até concluir a tarefa ou o trabalho.
4. Gosta de cooperar com os outros.
5. É original, tem sempre novas ideias.
6. É temperamental, muda de humor facilmente.
7. É inventivo, criativo.
8. É prestativo e ajuda os outros.
9. É amavel, tem consideração pelos outros.
10. Faz as coisas com eficiencia.
11. É sociavel, extrovertido.
12. É cheio de energia.
13. É um trabalhador de confiança.
14. Tem uma imaginacao fértil.

15. Fica tenso com frequência.
16. Fica nervoso facilmente.
17. Gera muito entusiasmo.
18. Gosta de refletir, brincar com as ideias.
19. Tem capacidade de perdoar, perdoa fácil.
20. Preocupa-se muito com tudo.

ANEXO B CRONOGRAMA

B.1 Principais etapas da dissertação

Tabela 14: Cronograma de atividades da dissertação

Atividade	**2020**						**2021**		
	Jul	**Ago**	**Set**	**Out**	**Nov**	**Dez**	**Jan**	**Fev**	**Mar-Jun**
Validação do Sistema de Coleta de Traços de Personalidade	x	x							
F1: Coleta de Dados de Personalidade						x	x	x	x
F1: Extração de Pegadas Ativas							x	x	x
F1: Extração de Pegadas Passivas							x	x	x
F1: Extração de Dados Demográficos						x	x	x	x
F1: Segmentação Inicial dos Grupos de Usuários								x	
F2: Otimização dos Agrupamentos Gerados								x	x
F2: Análise Descritiva dos Dados Agrupamentos Gerados							x	x	x
Redação da Dissertação	x	x	x	x	x	x	x	x	x
Entrega da Dissertação									x
Defesa da Dissertação									x

ANEXO C DADOS

C.1 Síntese do conjunto de dados unificado da pesquisa com todas as dimensões

Tabela 15: Listagem de todas colunas presentes no conjunto de dados

Dimensão	Tipo	Registros	Categoria	Tipo de pegada digital	Descrição
user_id	int64	188	Identificador	Não se aplica	Idetnficador de cada usuário na rede social Wedy
event_date	object	183	Demografico	Ativa	Data do evento
signed_up	object	188	Demografico	Ativa	Data do cadastro
gender	object	179	Demografico	Ativa	Genêro que o usuário se identifica
state	object	143	Demografico	Passiva	Estado de origem do usuário
event_budget	float64	188	Demografico	Ativa	Investimento no evento
event_style	float64	188	Demografico	Ativa	Estilo do evento
event_type	float64	188	Demografico	Ativa	Tipo do evento
event_size	float64	188	Demografico	Ativa	Tamanho do evento
extroversion	float64	188	Personalidade	Ativa	Traço de personalidade de extroversão
conscientiousness	float64	188	Personalidade	Ativa	Traço de personalidade de concensiosidade
agreeableness	float64	188	Personalidade	Ativa	Traço de personalidade de amabilidade
openness	float64	188	Personalidade	Ativa	Traço de personalidade de abertura à experiências
neuroticism	float64	188	Personalidade	Ativa	Traço de personalidade de neuroticismo
total_visits	int64	188	Comportamental	Passiva	Total de visitas realizadas em conteúdos da rede social
total_publications	int64	188	Comportamental	Ativa	Total de publicações realizadas na rede social
total_likes	int64	188	Comportamental	Ativa	Total de curtidas deixadas na rede social
total_comments	int64	188	Comportamental	Ativa	Total de comentários deixados na rede
visits_received	float64	188	Comportamental	Passiva	Total de visitas recebidas por outros usuários
likes_received	float64	188	Comportamental	Ativa	Total de curtidas recebidas de outros usuários
comments_received	float64	188	Comportamental	Ativa	Total de comentários recebidos de outros usuários
event_planning_dur_mths	float64	188	Comportamental	Ativa	Tempo planejando o casamento
gifts_received	float64	188	Comportamental	Passiva	Presentes recebidos
guests_vists	int64	188	Comportamental	Passiva	Visitas de convidados do evento
tasks_done	int64	188	Comportamental	Ativa	Tarefas de organização do evento concluídas
Marítimo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Palavra	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
#Madrinha	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
100	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
15 Anos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
160	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
160 G	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
160 M	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
1ª Vez	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
2020	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
36/38	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
38/40	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
40/42	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
42/44	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
90 Cm	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
90cm	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
A Gente	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
A Mão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
A Noiva	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
A Pena	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Abraço	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Acessório	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Acessórios	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Adolescente	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Agora	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Alfabeto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Alguém	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Alimento	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Almofada	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Alpendre	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Altura	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Amarelo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Amor	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Anel	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Animal	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Anágua	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Anúncio	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Ao Ar Livre	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Apenas 1 Vez	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas

Apenas Uma Vez	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Arbour	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Arecaceae	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Areia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Arquitetura	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Arranjo De Flores	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Arranjo Flor	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Arte	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Assoalho	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Atividades De Lazer	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Automóvel	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Azul	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Barba	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Barra	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bebida	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bebê	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Beijando	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Beijar	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Beijo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Biquíni	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Blazer	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Blusa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Boca	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bojo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bola	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bolo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bolo De Aniversário	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bolo De Casamento	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bolsa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Boné	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bordado	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bordados	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Bosque	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Botões	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Branco	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Brasil	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Brinquedo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Brochura	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Buquê De Flores	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Busto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cabelo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cabeça	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cada Detalhe	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cadeira	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Caia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Caixa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Calda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Caligrafia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Calça	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Calçado	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Calçados	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Calças	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Calções	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Caminho	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Camisa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Camiseta	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Campo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cara	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Carro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cartaz	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cartão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cartão De Saudação	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Casa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Casaco	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Casamento Civil	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Casar	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cauda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cerveja	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cerâmica	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cetim	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Chapéu	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Chão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cidade	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cinto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cintura	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Colagem	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Colar	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas

Comida	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Comprimento	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Confeitaria	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Construção	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Contador De Barras	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Convite	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Convites	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cor	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Coração	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Corpo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Correio	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Corrimão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Costa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Costas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Creme	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Criança	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Cristais	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Céu	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Decoração	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Decoração De Casa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Decote	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Decote Nas Costas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Dedo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Dentes	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Dentro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Dentro De Casa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Design De Interiores	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Design Floral	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Detalhes	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Deus	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Dia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Doces	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Documento	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Edifício	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Ela	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Ele	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Eletrônicos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Envelope	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Ervas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Espaço	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Esporte	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Esportes	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Esse	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Estrada	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Estreia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Eua	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Face	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Fake Out	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Família	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Fantasia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Fato	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Felicidade	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Feminino	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Festa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Flor	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Flores	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Florescer	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Floresta	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Florestal	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Folha	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Folheto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Formatura	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Forro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Foto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Fotografia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Francesa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Francês	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Frente	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Férias	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Garota	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Garrafa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Gente	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Glacê	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Grama	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Gramado	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Grande Dia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Gravata	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas

Gráficos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Hoje	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Homem	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Igreja	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Iluminação	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Interior	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Iorane	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Iquebana	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Jade	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Jaqueta	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Jarda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Jardim	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Jarra	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Jeans	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Jóias	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Kid	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Lado Costeiro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Laranja	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Lavanderia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Leve	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Lindo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Linho	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Livreto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Livro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Loira	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Longo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Luz Do Sol	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Lábio	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Madeira	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Madeira Compensada	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Madrinhas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mamífero	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Manga	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Manga Comprida	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Manga Longa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mar	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Medida	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Medidas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Menina	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mensagem	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Menu	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mesa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mesa De Jantar	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mil	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mim	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Miúdo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mobília	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Moda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Modelo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Modelo Sereia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Moradia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Muito Amor	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Muito Feliz	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mulher	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Multidão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Máquina	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mãe	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mãe Da Noiva	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mãe De Noiva	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Mão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Móveis	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Nada	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Namorando	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Namoro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Nas Costas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Natureza	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
No Corpo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Noite	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Noiva	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Noiva Branco	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Noivado	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Noivas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Noivo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Noivos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Nude	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Nós	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Número	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas

O Amor	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
O Casamento	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
O Dia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
O Grande Dia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
O Vestido	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
O Véu	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Oceano	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Off	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Off White	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Ornamento	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Padrinhos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Padrão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Palmeira	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Panfleto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Papel	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Parque	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pedrarias	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pele	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Perfeito Estado	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Planta	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Planta Em Vasos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Plantador	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Plantar	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Poa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Porta	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pose De Dança	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Praia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Prateleira	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Prato	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Presente	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Princesa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Propaganda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pré Wedding	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pub	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Página	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pássaro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pérolas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Pétala	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Quadril	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Quarto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Quem	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Quintal	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Recomendação	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Refeição	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Renda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Renda Francesa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Restaurante	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Retrato	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Rindo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Robe	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Rocha	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Roda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Romance	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Rosa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Rosto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Roupa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Roupa De Banho	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Roupas	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Roupão	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Roxo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Rural	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Réveillon	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Rótulo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Saia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sala	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sala De Estar	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Salmao	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sandália	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sapato	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Save The Date	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Seda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Selfie	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Selva	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sentado	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sentando	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sereia	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Short	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas

Shorts	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sinalizador	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Smoking	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sobremesa	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sobretudo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sofá	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Somente Uma Vez	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sonho	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sonhos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sorria	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Sorriso	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Swarovski	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Símbolo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tamanho	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tamanho 36	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tamanho 38	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tamanho 40	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tamanho 42	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tamanho M	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tamanho P	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tatuagem	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tecido	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tela	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Telefone	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Terno	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Terra	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Texto	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tiffany	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Todo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Todo Bordado	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Todo Em Renda	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Todos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tomara	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Traje	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Transparência	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Transporte	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Travesseiro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tudo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Tule	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Uma Vez	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Uma única Vez	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Urbano	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vaso	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vegetação	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Verde	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vermelho	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Veste	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestido	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestido Branco	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestido De Casamento	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestido De Noite	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestido De Noiva	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestido Longo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestidos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestimenta	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vestuário	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Veículo	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vida Noturna	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Vidro	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Você	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Véu	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Wedy	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Zara	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Zíper	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Água	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Álcool	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Árvore	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Óculos	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
Óculos De Sol	float64	153	Comportamental	Passiva	Tópico de interesse expresso pelo número de visitas
monthly_retention	float64	6	Comportamental	Passiva	Tópico de interesseexpresso pelo número de visitas
published_content_text	object	35	Comportamental	Passiva	Tópico de interesseexpresso pelo número de visitas
published_content_image	object	31	Comportamental	Passiva	Tópico de interesseexpresso pelo número de visitas

C.2 Distribuição de valores de cada dimensão do conjunto Principais Dimensões em seus dois agrupamentos

		Cluster 1	Cluster 2
total_visits	*valor médio*	16.719101	7.865672
	valor mínimo	3.000000	3.000000
	valor máximo	147.000000	34.000000
total_publications	*valor médio*	0.539326	0.477612
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	9.000000
total_likes	*valor médio*	2.831461	0.686567
	valor mínimo	0.000000	0.000000
	valor máximo	66.000000	20.000000
visits_received	*valor médio*	4.505618	5.820896
	valor mínimo	0.000000	0.000000
	valor máximo	79.000000	73.000000
likes_received	*valor médio*	1.191011	1.029851
	valor mínimo	0.000000	0.000000
	valor máximo	20.000000	12.000000
state	*valor médio*	11.808989	10.597015
	valor mínimo	0.000000	0.000000
	valor máximo	17.000000	17.000000
event_budget	*valor médio*	1.280899	-1.000000
	valor mínimo	-1.000000	-1.000000
	valor máximo	3.000000	-1.000000
event_style	*valor médio*	6.988764	-1.000000
	valor mínimo	-1.000000	-1.000000
	valor máximo	9.000000	-1.000000
event_type	*valor médio*	3.000000	-0.402985
	valor mínimo	3.000000	-1.000000
	valor máximo	3.000000	3.000000
event_size	*valor médio*	1.348315	-0.970149
	valor mínimo	-1.000000	-1.000000
	valor máximo	3.000000	1.000000
event_planning_duration_months	*valor médio*	15.555807	8.584573
	valor mínimo	0.800000	1.300000
	valor máximo	69.300000	26.700000
gifts_received	*valor médio*	1.337079	0.761194
	valor mínimo	0.000000	0.000000
	valor máximo	38.000000	14.000000
guests_vists	*valor médio*	145.932584	75.985075
	valor mínimo	0.000000	0.000000
	valor máximo	1739.000000	945.000000
extroversion	*valor médio*	2.471910	2.343284
	valor mínimo	1.000000	1.000000
	valor máximo	4.000000	4.000000
conscientiousness	*valor médio*	2.348315	2.089552
	valor mínimo	1.000000	1.000000
	valor máximo	4.000000	4.000000
agreeableness	*valor médio*	2.314607	2.119403
	valor mínimo	1.000000	1.000000
	valor máximo	4.000000	4.000000
openness	*valor médio*	2.314607	2.268657
	valor mínimo	1.000000	1.000000
	valor máximo	4.000000	4.000000
neuroticism	*valor médio*	2.314607	2.507463
	valor mínimo	1.000000	1.000000
	valor máximo	4.000000	4.000000
Acessório	*valor médio*	0.348315	0.313433
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	6.000000
Acessórios	*valor médio*	0.359551	0.313433
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	6.000000
Alguém	*valor médio*	0.258427	0.014925
	valor mínimo	0.000000	0.000000
	valor máximo	10.000000	1.000000
Amarelo	*valor médio*	1.842697	0.835821
	valor mínimo	0.000000	0.000000
	valor máximo	35.000000	7.000000
Amor	*valor médio*	0.764045	0.522388
	valor mínimo	0.000000	0.000000
	valor máximo	9.000000	7.000000
Anúncio	*valor médio*	0.337079	0.358209
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	13.000000
Ao Ar Livre	*valor médio*	0.528090	0.373134
	valor mínimo	0.000000	0.000000
	valor máximo	10.000000	4.000000
Branco	*valor médio*	1.359551	0.626866

		Cluster 1	Cluster 2
	valor mínimo	0.000000	0.000000
	valor máximo	25.000000	7.000000
Brochura	*valor médio*	0.303371	0.298507
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	10.000000
Calçado	*valor médio*	0.224719	0.119403
	valor mínimo	0.000000	0.000000
	valor máximo	4.000000	2.000000
Cartaz	*valor médio*	0.550562	0.328358
	valor mínimo	0.000000	0.000000
	valor máximo	12.000000	13.000000
Casaco	*valor médio*	0.325843	0.164179
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	6.000000
Casar	*valor médio*	0.438202	0.223881
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	2.000000
Convite	*valor médio*	2.033708	1.000000
	valor mínimo	0.000000	0.000000
	valor máximo	33.000000	9.000000
Convites	*valor médio*	2.044944	0.910448
	valor mínimo	0.000000	0.000000
	valor máximo	23.000000	12.000000
Deus	*valor médio*	0.730337	0.611940
	valor mínimo	0.000000	0.000000
	valor máximo	8.000000	7.000000
Ele	*valor médio*	2.134831	1.089552
	valor mínimo	0.000000	0.000000
	valor máximo	38.000000	7.000000
Face	*valor médio*	0.539326	0.119403
	valor mínimo	0.000000	0.000000
	valor máximo	11.000000	3.000000
Família	*valor médio*	0.202247	0.104478
	valor mínimo	0.000000	0.000000
	valor máximo	4.000000	3.000000
Felicidade	*valor médio*	0.157303	0.149254
	valor mínimo	0.000000	0.000000
	valor máximo	2.000000	4.000000
Feminino	*valor médio*	0.887640	0.283582
	valor mínimo	0.000000	0.000000
	valor máximo	13.000000	4.000000
Flor	*valor médio*	0.505618	0.164179
	valor mínimo	0.000000	0.000000
	valor máximo	7.000000	4.000000
Folheto	*valor médio*	0.269663	0.298507
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	10.000000
Grama	*valor médio*	0.168539	0.014925
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	1.000000
Grande Dia	*valor médio*	0.438202	0.164179
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	3.000000
Laranja	*valor médio*	0.382022	0.283582
	valor mínimo	0.000000	0.000000
	valor máximo	3.000000	2.000000
Moda	*valor médio*	0.921348	0.447761
	valor mínimo	0.000000	0.000000
	valor máximo	11.000000	5.000000
Natureza	*valor médio*	0.337079	0.283582
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	4.000000
Noiva	*valor médio*	0.685393	0.283582
	valor mínimo	0.000000	0.000000
	valor máximo	21.000000	4.000000
Noivo	*valor médio*	0.202247	0.104478
	valor mínimo	0.000000	0.000000
	valor máximo	3.000000	3.000000
Nós	*valor médio*	0.865169	0.522388
	valor mínimo	0.000000	0.000000
	valor máximo	9.000000	4.000000
O Grande Dia	*valor médio*	0.359551	0.029851
	valor mínimo	0.000000	0.000000
	valor máximo	8.000000	1.000000
Papel	*valor médio*	0.528090	0.313433
	valor mínimo	0.000000	0.000000

		Cluster 1	Cluster 2
	valor máximo	12.000000	10.000000
Planta	*valor médio*	0.842697	0.432836
	valor mínimo	0.000000	0.000000
	valor máximo	13.000000	7.000000
Plantar	*valor médio*	0.629213	0.104478
	valor mínimo	0.000000	0.000000
	valor máximo	10.000000	2.000000
Robe	*valor médio*	0.505618	0.328358
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	5.000000
Rosa	*valor médio*	0.415730	0.164179
	valor mínimo	0.000000	0.000000
	valor máximo	19.000000	7.000000
Rosto	*valor médio*	0.595506	0.417910
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	5.000000
Roupa	*valor médio*	0.146067	0.014925
	valor mínimo	0.000000	0.000000
	valor máximo	3.000000	1.000000
Roupas	*valor médio*	1.741573	0.283582
	valor mínimo	0.000000	0.000000
	valor máximo	35.000000	5.000000
Roupão	*valor médio*	0.393258	0.104478
	valor mínimo	0.000000	0.000000
	valor máximo	9.000000	3.000000
Salmao	*valor médio*	1.011236	0.447761
	valor mínimo	0.000000	0.000000
	valor máximo	12.000000	7.000000
Sapato	*valor médio*	0.404494	0.149254
	valor mínimo	0.000000	0.000000
	valor máximo	8.000000	3.000000
Save The Date	*valor médio*	0.752809	0.208955
	valor mínimo	0.000000	0.000000
	valor máximo	15.000000	4.000000
Sonhos	*valor médio*	0.179775	0.104478
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	2.000000
Sorria	*valor médio*	0.359551	0.119403
	valor mínimo	0.000000	0.000000
	valor máximo	6.000000	3.000000
Texto	*valor médio*	1.337079	0.402985
	valor mínimo	0.000000	0.000000
	valor máximo	17.000000	6.000000
Todos	*valor médio*	0.258427	0.179104
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	6.000000
Tudo	*valor médio*	0.752809	0.462687
	valor mínimo	0.000000	0.000000
	valor máximo	10.000000	4.000000
Vermelho	*valor médio*	1.617978	0.776119
	valor mínimo	0.000000	0.000000
	valor máximo	25.000000	7.000000
Vestido	*valor médio*	1.123596	0.343284
	valor mínimo	0.000000	0.000000
	valor máximo	43.000000	4.000000
Vestido De Noite	*valor médio*	0.438202	0.208955
	valor mínimo	0.000000	0.000000
	valor máximo	7.000000	3.000000
Vestido Longo	*valor médio*	0.797753	0.417910
	valor mínimo	0.000000	0.000000
	valor máximo	9.000000	5.000000
Vestidos	*valor médio*	0.707865	0.119403
	valor mínimo	0.000000	0.000000
	valor máximo	38.000000	5.000000
Vestimenta	*valor médio*	0.146067	0.014925
	valor mínimo	0.000000	0.000000
	valor máximo	3.000000	1.000000
Vestuário	*valor médio*	2.808989	1.104478
	valor mínimo	0.000000	0.000000
	valor máximo	41.000000	8.000000
Você	*valor médio*	0.516854	0.298507
	valor mínimo	0.000000	0.000000
	valor máximo	5.000000	6.000000
Árvore	*valor médio*	0.359551	0.208955
	valor mínimo	0.000000	0.000000
	valor máximo	8.000000	4.000000

		Cluster 1	Cluster 2
event_date_season	*valor médio*	**2.584270**	**2.134328**
	valor mínimo	**1.000000**	**-1.000000**
	valor máximo	**4.000000**	**4.000000**
event_date_weekday	*valor médio*	**4.786517**	**4.134328**
	valor mínimo	**0.000000**	**-1.000000**
	valor máximo	**6.000000**	**6.000000**
signed_up_weekday	*valor médio*	**2.707865**	**2.880597**
	valor mínimo	**0.000000**	**-1.000000**
	valor máximo	**6.000000**	**6.000000**
published_content_text_word_count	*valor médio*	**5.056180**	**4.671642**
	valor mínimo	**1.000000**	**1.000000**
	valor máximo	**76.000000**	**70.000000**
published_content_text_char_count	*valor médio*	**27.123596**	**21.910448**
	valor mínimo	**0.000000**	**0.000000**
	valor máximo	**433.000000**	**415.000000**
published_content_cogmech_count	*valor médio*	**1.393258**	**1.447761**
	valor mínimo	**0.000000**	**0.000000**
	valor máximo	**26.000000**	**30.000000**
published_content_affect_count	*valor médio*	**0.460674**	**0.477612**
	valor mínimo	**0.000000**	**0.000000**
	valor máximo	**12.000000**	**14.000000**
published_content_social_count	*valor médio*	**0.685393**	**0.656716**
	valor mínimo	**0.000000**	**0.000000**
	valor máximo	**11.000000**	**21.000000**
published_content_verb_count	*valor médio*	**0.494382**	**0.537313**
	valor mínimo	**0.000000**	**0.000000**
	valor máximo	**8.000000**	**10.000000**

C.3 Distribuição de valores de cada dimensão do conjunto Traços de Personalidade & Principais índices de comportamento em seus dois agrupamentos

		Cluster 1	Cluster 2
total_publications	mean	0.456522	0.600000
	min	0.000000	0.000000
	max	7.000000	9.000000
	std	1.180548	1.518634
	mode	0.000000	0.000000
total_comments	mean	0.054348	0.138462
	min	0.000000	0.000000
	max	2.000000	4.000000
	std	0.271913	0.583013
	mode	0.000000	0.000000
comments_received	mean	0.010870	0.092308
	min	0.000000	0.000000
	max	1.000000	3.000000
	std	0.104257	0.458362
	mode	0.000000	0.000000
likes_received	mean	0.967391	1.353846
	min	0.000000	0.000000
	max	12.000000	20.000000
	std	2.442531	3.572599
	mode	0.000000	0.000000
visits_received	mean	4.673913	5.769231
	min	0.000000	0.000000
	max	61.000000	79.000000
	std	12.466314	15.881169
	mode	0.000000	0.000000
total_likes	mean	2.163043	1.523077
	min	0.000000	0.000000
	max	66.000000	27.000000
	std	8.611616	4.426863
	mode	0.000000	0.000000
openness	mean	1.695652	3.153846
	min	1.000000	1.000000
	max	4.000000	4.000000
	std	0.946238	0.905379
	mode	1.000000	4.000000
extroversion	mean	1.793478	3.307692
	min	1.000000	1.000000
	max	4.000000	4.000000
	std	0.805722	0.789048
	mode	1.000000	4.000000
agreeableness	mean	1.826087	2.830769
	min	1.000000	1.000000
	max	4.000000	4.000000
	std	0.872141	0.893890

		Cluster 1	Cluster 2
	mode	1.000000	3.000000
total_visits	mean	12.282609	16.861538
	min	3.000000	3.000000
	max	99.000000	211.000000
	std	16.620738	33.004203
	mode	5.000000	5.000000
neuroticism	mean	2.239130	2.630769
	min	1.000000	1.000000
	max	4.000000	4.000000
	std	1.031052	1.206353
	mode	2.000000	4.000000
conscientiousness	mean	1.663043	3.046154
	min	1.000000	1.000000
	max	4.000000	4.000000
	std	0.867954	0.925857
	mode	1.000000	3.000000