Revista Brasileira de Computação Aplicada, Julho, 2022

DOI: 10.5335/rbca.v14i2.12755
Vol. 14, Nº 2, pp. 1–15
Homepage: seer.upf.br/index.php/rbca/index

ARTIGO ORIGINAL

Agrupando e analisando o comportamento de usuários de redes sociais a partir da combinação de traços de personalidade, dados demográficos e pegadas digitais

Clustering and analyzing social network users behavior by combining personality traits and digital footprints

Daniel Tamiosso[1] and Prof. Dra. Patricia A. Jaques[1]

[1]Universidade do Vale do Rio dos Sinos

*danieltamiosso@gmail.com; pjaques@unisinos.br

Recebido: 18/07/2021. **Revisado:** 18/04/2022. **Aceito:** 04/05/2022.

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 e passivas, quando produzidas de forma não intencional. Este trabalho busca identificar perfis de usuários em redes sociais a partir do agrupamento de dados de comportamento em redes sociais, dados demográficos e informações socioafetivas. Dessa forma, verifica-se a viabilidade na criação de grupos significativos, 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, foram empregados algoritmos de aprendizado não supervisionados (clusterização). Embora esse trabalho analise um grupo pequeno de usuários (157 participantes), pode-se verificar 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 campos subjacentes relacionados.

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

Abstract

Digital social networks are becoming more and more popular, offering a massive platform for analyzing human behavior in computer-mediated contexts. Human behavior can be explored by analyzing the set of digital footprints left by people when interacting with social networks. Digital footprints can be classified into active and passive when produced unintentionally. This work seeks to identify user profiles in social networks from the grouping of behavior data in social networks, demographic data, and socio-affective information. Thus, the feasibility of creating meaningful groups is verified, as well as a qualitative and quantitative analysis of the groups produced is made available, in order to understand the quality of the groups formed and their validity in relation to the revised knowledge of personality psychology. More specifically, unsupervised learning algorithms (clustering) were employed. Although this work analyzes a small group of users (157 participants), correlations observed in the related bibliography can be verified, being the first step for future proposals in order to raise awareness about the relationship of social networks, personality computation, and its related fields.

Keywords: Big Five Personality Traits; Clustering; Digital Footprints; Personality Computing; Social Networks

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 and 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 and 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 et al., 2015).

Essa também é uma oportunidade para muitos setores comerciais, da publicidade, assim como para o desenvolvimento de produtos digitais (Chen et al., 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 et al., 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 and Griffiths, 2011) – para fornecer mensagens de marketing altamente personalizadas e ajustar seus produtos para melhor adequação ao perfil psicológico de cada usuário (Youyou et al., 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 et al., 2015). Nesse contexto, os trabalhos que visam a detecção de traços de personalidade utilizam essencialmente a abordagem supervisionada. Entretanto, essa pesquisa 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 et al., 2015), 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.

Este trabalho 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 detalhados 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, que possui uma rede social especializada no planejamento de eventos de casamento, foi realizada a coleta permissiva 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), totalizando 157 participantes selecionados, em um conjunto de dados com aproximadamente 450 dimensões.

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 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 (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.

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 and 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. Ela 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.

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.

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 and 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 and Mohammadi, 2014).

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

- **Abertura à Experiência**: A abertura está relacionada à imaginação, criatividade, curiosidade, tolerância, liberalismo político e apreciação pela cultura;
- **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;
- **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;
- **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;
- **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;

3.1 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 and Kosinski, 2014). 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 and Macy, 2014).

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).

3.1.1 *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 and 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.

3.1.2 *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 and 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.

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 and Pennington, 2018).

4 Mineração de dados e algoritmos de clusterização

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).

Como esse trabalho 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 and Liu, 2011). (Dutt et al., 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".

5 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, 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.

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 and 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 and Elkan (2004), a correta escolha de k fica mais difícil quando os dados têm muitas dimensões, mesmo quando os clusters estão bem separados. 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 and Wu (2012), Celebi et al. (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.
- **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 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 clusterização tradicionais. O algoritmo baseia-se nos princípios de vetores de Eigen de alguma matriz com base na distância entre os pontos (ou outras propriedades) e, em seguida, usá-os para agrupar os vários pontos (Verma and 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. 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.

6 Trabalhos relacionados

Os trabalhos analisados envolveram diferentes vieses de estudos de traços de personalidades e sua aplicabilidade em redes sociais. A revisão dos trabalhos relacionados mostram como a personalidade de um usuário está intrinsecamente conectada as suas preferências e características de navegação em redes sociais (Stillwell and Kosinski, 2004,

[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.

Kosinski et al., 2015, 2014, Burbach et al., 2019). 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. 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 (Lambiotte and Kosinski, 2014, Muhammad et al., 2018, Azucar et al., 2018, Segalin et al., 2017). 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, sendo aquele que possui maior concentração de pesquisas e trabalhos relacionados, alcançando resultados de julgamento de personalidade melhores do que o próprio julgamento humano (Kaushal and Patwardhan, 2018, Marengo and Settanni, 2019, Azucar et al., 2018, Youyou et al., 2015). Embora essa área de pesquisa enfrente desafios pela ausência de datasets públicos para treinamento e validação de suas hipóteses. 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 (Halim et al., 2019, Kosinski et al., 2013, Wald et al., 2012, Adali and Golbeck, 2012, Golbeck, Robles and Turner, 2011, Golbeck, Robles, Edmondson and Turner, 2011, Ortigosa et al., 2014, Markovikj et al., 2021). 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 (Kaushal and Patwardhan, 2018, Chin and Wright, 2014, Golbeck, Robles, Edmondson and Turner, 2011, Farnadi et al., 2013, Appling et al., 2013, Coltheart, 1981, Adali and Golbeck, 2012). Dado todas essas condições e o estado-da-arte dessa área de conhecimento, conclui-se que a detecção da personalidade é uma ferramenta que influencia o comportamento dos usuários nas redes sociais (Kaushal and Patwardhan, 2018, Matz and Kosinski, 2019, Matz et al., 2017, Chen et al., 2019). 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 (Matz et al., 2020, Smith, 2018, Tikkinen-Piri et al., 2018, Maldonado et al., 2019, Isaak and Hanna, 2018). 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 e o quão relevantes eles são na segmentação dos usuários quando combinados com pegadas digitais.

7 Metodologia

Para o desenvolvimento do trabalho e alcance dos seus objetivos, um conjunto de dados de alta-dimensionalidade, contendo dados comportamentais e demográficos da rede social da Wedy [3]. 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.

As seguintes etapas foram conduzidas para a realização da pesquisa:

- 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;
- 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;
- Combinação dos dados de personalidade e comportamento com as informações demográficas cadastradas por esses usuários na rede social;
- 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 - o ambiente computacional utilizado foi composto da linguagem de programação Python e bibliotecas de Aprendizado de Máquina como *Sklearn* e *Scipy*, bem como *Seaborn* e *Plotly* para visualização dos dados;
- 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;
- 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;

7.1 Dados

Os dados disponibilizados pela Wedy são separados em três matrizes relativamente esparsas contendo aproximadamente 450 colunas, com a combinação de ambos. 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.

O conjunto de dados oferecidos para a pesquisa potencialmente contém uma série de inconsistências e redundâncias, sendo na sua essência imperfeitos para o de-

[3] A Wedy é uma startup de organização de casamentos que possui uma funcionalidade que conecta seus usuários em uma comunidade online

senvolvimento imediato desse trabalho. 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.

A partir da conclusão de coleta de dados e do pré-processamento de dados, busca-se por 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 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, 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.

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 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 et al., 2017).

8 Análise e resultados

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. 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), 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[4] na rede social, ao efetuarem o acesso ao produto no seguinte endereço https://app.wedy.com/quiz.

Porém, devido às limitações impostas pela pandemia da Covid-19, e a relação da rede social com eventos presenciais, onde foram restringidos das suas realizações desde esse momento até, no mínimo, a conclusão da pesquisa (Gössling et al., 2020), apenas 285 usuários tiveram seus traços de personalizados obtidos pelo formulário online desenvolvido por essa pesquisa.

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, 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).

Alguns dados do conjunto de dados podem ser definitivamente *outliers*[5], 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

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

[5] 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).

desvios padrões do conjunto de dados de trabalho. Nos dados disponibilizados pela rede social, há alguns casos em que elementos particulares estão ausente. 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. 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 o conjunto de dados possui tipos mistos de dados: categóricos e discretos, todos os conjuntos foram discretizados e posteriormente escalonados.

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, 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. Também foram escolhidas três estratégias de seleção de características:

- **Todas Dimensões**: seleção que compreende todas as características mantidas após o processo de pré-processamento. 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.
- **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.

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 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.

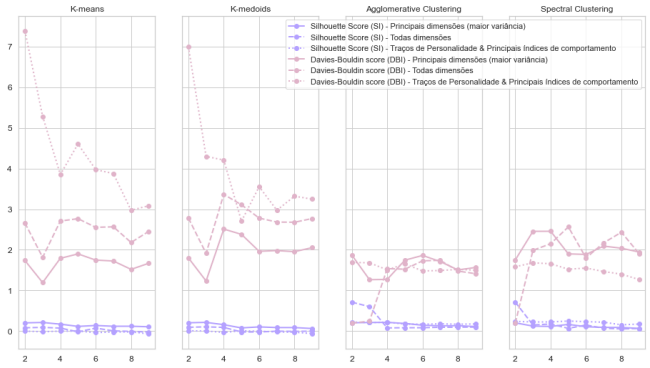

Figura 1: Os valores DBI e SC para as várias formações de agrupamento do conjunto de dados usando K-means, K-medoides, Agglomerative Clustering e Spectral Clustering

8.1 Análise descritiva dos agrupamentos gerados

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

Utilizando uma abordagem supervisionada, pode-se avaliar a contribuição das variáveis presentes em cada um dos conjuntos de dados na formação dos agrupamento (Ismaili et al., 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 Fig. 2, 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 (Fig. 3).

Seguindo a análise descritiva, baseada na formação dos dois agrupamentos criados a partir do conjunto de dados denominado Principais Dimensões, pode-se observar nas Figs. 7 a 11 a distribuição dos valores escalonados entre os clusters. 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, 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 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).

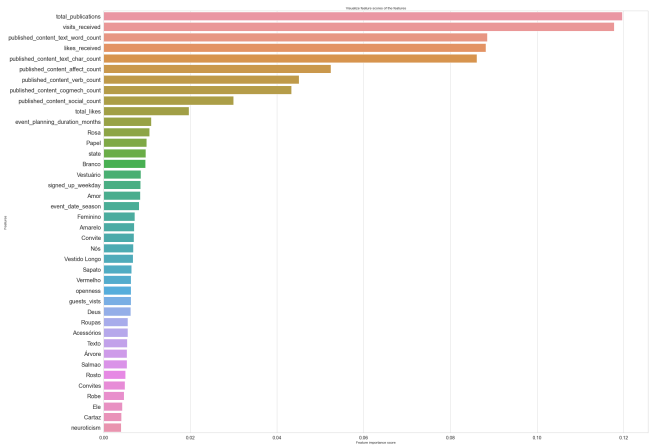

Figura 2: Coeficiente de importância de cada dimensão na formação dos agrupamentos para as Principais Dimensões

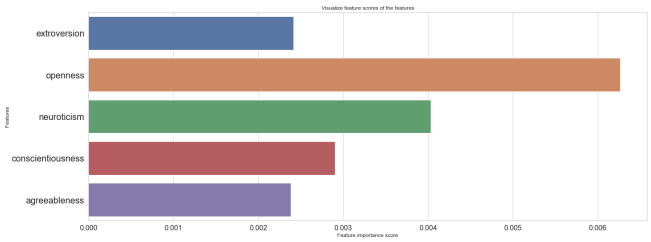

Figura 3: Coeficiente de importância dos traços de personalidade na formação dos agrupamentos para as Principais Dimensões

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 comportamen-

tal 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.

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 e 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 Fig. 5 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 Fig. 5, agrupa os usuários que possuem índices mais altos de abertura à experiência, amabilidade, conscienciosidade e, principalmente, extroversão, com uma pontuação relativa inferior de neuroticismo.

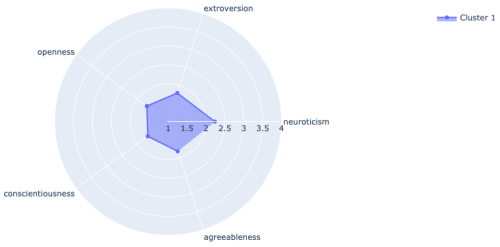

Figura 4: Distribuição dos traços de personalidade no modelo dos Cinco Grande Fatores para o Cluster 1

Analisando a distribuição das dimensões nos dois clusters, observa-se visualmente na Fig. 6 com distribuição dos valores escalonados 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 neuroticismo 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).

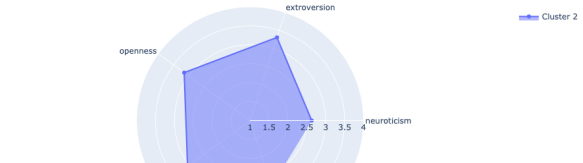

Figura 5: Distribuição dos traços de personalidade no modelo dos Cinco Grande Fatores para o Cluster 2

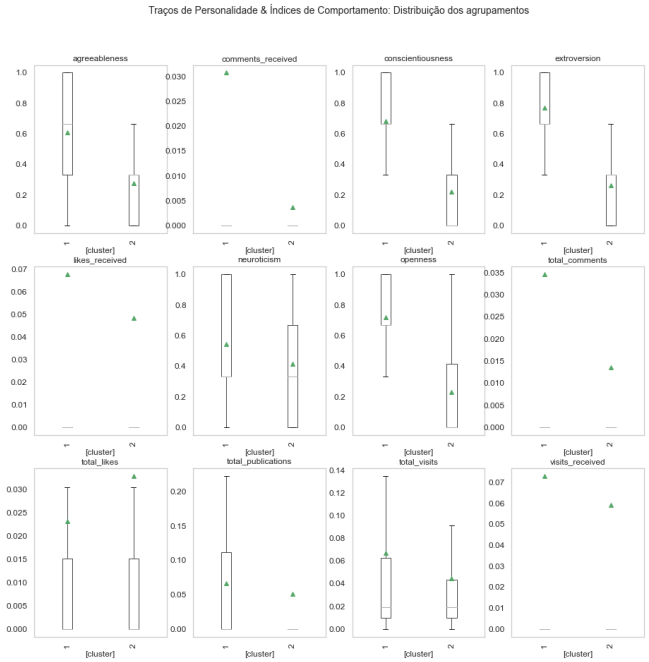

Figura 6: Traços de Personalidade e Índices de Comportamento: Distribuição dos agrupamentos

De forma sintética, agrupou-se os indicadores que foram mais significativos em cada agrupamento. Com isso, observou-se que o Cluster 1 possui um engajamento significativo na rede social, embora menor em relação ao Clus-

ter 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.

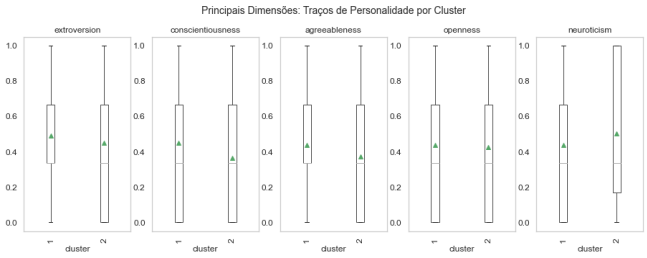

Figura 7: Análise de distribuição dos clusters no conjunto Principais Dimensões: Traços de Personalidade

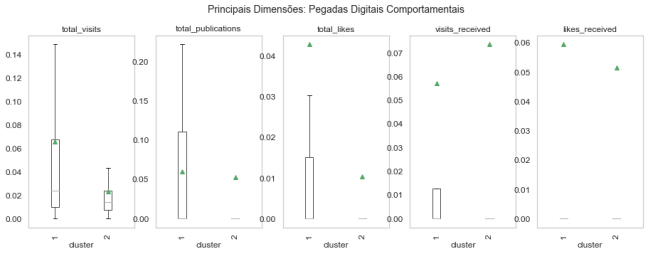

Figura 8: Análise de distribuição dos clusters no conjunto Principais Dimensões: Pegadas Digitais Comportamentais

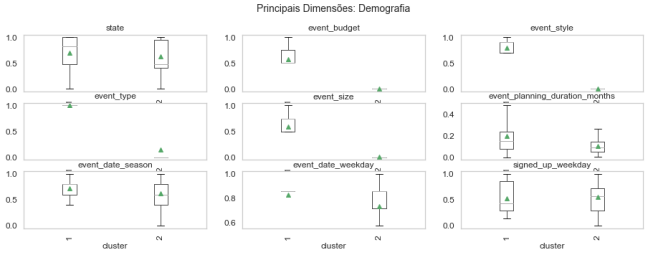

Figura 9: Análise de distribuição dos clusters no conjunto Principais Dimensões: Demografia

9 Discussão

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

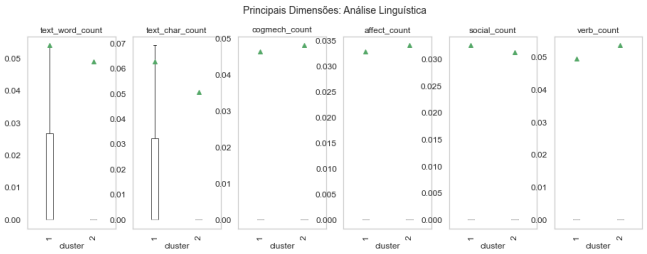

Figura 10: Análise de distribuição dos clusters no conjunto Principais Dimensões: Análise Linguística

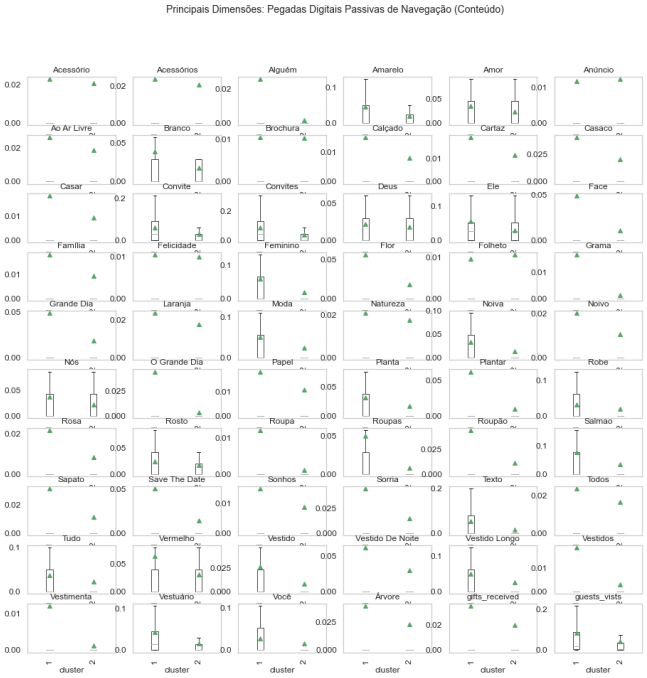

Figura 11: Análise de distribuição dos clusters no conjunto Principais Dimensões: Pegadas Digitais Passivas de Comportamento (categoria de conteúdo)

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 estraté-

gicas 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 que indica a preferência por uma abordagem organizada da vida, em contraste com uma abordagem espontânea.

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 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 and 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 and 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 observado 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.

10 Conclusão

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.

Referências

Ackermann, M. R., Blömer, J., Kuntze, D. and Sohler, C. (2014). Analysis of lomerative clustering, *Algorithmica* **69**(1): 184–215. https://doi.org/10.1007/s00453-012-9717-4.

Adali, S. and Golbeck, J. (2012). Predicting personality with social behavior, *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, IEEE, pp. 302–309. https://doi.org/10.1109/ASONAM.2012.58.

Aggarwal, C. C. (2015). *Data mining: the textbook*, Springer. https://doi.org/10.1007/978-3-319-14142-8.

Appling, D., Briscoe, E., Hayes, H. and Mappus, R. (2013). Towards automated personality identification using speech acts, Vol. 7, pp. 10–13. Disponível em https://ojs.aaai.org/index.php/ICWSM/article/view/14469.

Arakerimath, A. R. and Gupta, P. K. (2015). Digital footprint: Pros, cons, and future, *International Journal of Latest Technology in Engineering, Management & Applied Science* **4**(10): 52–56. Disponível em https://ijltemas.in/DigitalLibrary/Vol.4Issue10/52-56.pdf.

Arora, P., Varshney, S. et al. (2016). Analysis of k-means and k-medoids algorithm for big data, *Procedia Computer Science* **78**: 507–512. https://doi.org/10.1016/j.procs.2016.02.095.

Azucar, D., Marengo, D. and Settanni, M. (2018). Predicting the big 5 personality traits from digital footprints on social media: A meta-analysis, *Personality and Individual Differences* **124**: 150–159. https://doi.org/10.1016/j.paid.2017.12.018.

Barbier, G. and Liu, H. (2011). Data mining in social media, *Social network data analytics*, Springer, pp. 327–352. https://doi.org/10.1007/978-1-4419-8462-3_12.

Burbach, L., Halbach, P., Ziefle, M. and Calero Valdez, A. (2019). Who shares fake news in online social networks?, *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pp. 234–242. https://doi.org/10.1145/3320435.3320456.

Celebi, M. E., Kingravi, H. A. and Vela, P. A. (2013). A comparative study of efficient initialization methods for the k-means clustering algorithm, *Expert systems with applications* **40**(1): 200–210. https://doi.org/10.1016/j.eswa.2012.07.021.

Chen, Y.-J., Chen, Y.-M., Hsu, Y.-J. and Wu, J.-H. (2019). Predicting consumers' decision-making styles by analyzing digital footprints on facebook, *International Journal of Information Technology & Decision Making* **18**(02): 601–627. https://doi.org/10.1142/S0219622019500019.

Chen, Y., Pavlov, D. and Canny, J. F. (2009). Large-scale behavioral targeting, *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, pp. 209–218. https://doi.org/10.1145/1557019.1557048.

Chin, D. N. and Wright, W. R. (2014). Social media sources for personality profiling, *UMAP Workshops*. Disponível em http://ceur-ws.org/Vol-1181/empire2014_paper_09.pdf.

Coltheart, M. (1981). The mrc psycholinguistic database, *The Quarterly Journal of Experimental Psychology Section A* **33**(4): 497–505. https://doi.org/10.1080%2F14640748108400805.

Damasio, A. R. (1994). Descartes' error: Emotion, rationality and the human brain, *New York: Putnam* pp. 1061–1070.

Dutt, A., Ismail, M. A. and Herawan, T. (2017). A systematic review on educational data mining, *Ieee Access* **5**: 15991–16005. https://doi.org/10.1109/ACCESS.2017.2654247.

Dwork, C. and Mulligan, D. K. (2013). It's not privacy, and it's not fair, *Stan. L. Rev. Online* **66**: 35. Disponível em https://www.stanfordlawreview.org/online/privacy-and-big-data-its-not-privacy-and-its-not-fair/.

Farnadi, G., Zoghbi, S., Moens, M.-F. and De Cock, M. (2013). Recognising personality traits using facebook status updates, Vol. 7, pp. 14–18. Disponível em https://ojs.aaai.org/index.php/ICWSM/article/view/14470.

Fish, T. (2009). *My Digital Footprint A two-sided digital business model where your privacy will be someone else's business!*, Futuretext.

Freitag, M. and Bauer, P. C. (2016). Personality traits and the propensity to trust friends and strangers, *The social science journal* **53**(4): 467–476. https://doi.org/10.1016/j.soscij.2015.12.002.

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J. M. and Herrera, F. (2016). Big data preprocessing: methods and prospects, *Big Data Analytics* **1**(1): 1–22. https://doi.org/10.1186/s41044-016-0014-0.

Gavrilova, M. L. (2018). Machine learning for social behavior understanding, *Proceedings of Computer Graphics International 2018*, pp. 247–252. https://doi.org/10.1145/3208159.3208187.

Golbeck, J., Robles, C., Edmondson, M. and Turner, K. (2011). Predicting personality from twitter, *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, IEEE, pp. 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33.

Golbeck, J., Robles, C. and Turner, K. (2011). Predicting personality with social media, *CHI'11 extended abstracts on human factors in computing systems*, pp. 253–262. https://doi.org/10.1145/1979742.1979614.

Golder, S. A. and Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research, *Annual Review of Sociology* **40**: 129–152. https://doi.org/10.1146/annurev-soc-071913-043145.

Gössling, S., Scott, D. and Hall, C. M. (2020). Pandemics, tourism and global change: a rapid assessment of covid-19, *Journal of Sustainable Tourism* **29**(1): 1–20. https://doi.org/10.1080/09669582.2020.1758708.

Granville, K. (2018). Facebook and cambridge analytica: What you need to know as fallout widens, *The New York Times* **19**. Disponível em https://www.nytimes.com/2018/03/19/technology/facebook-cambridge-analytica-explained.html.

Halim, Z., Atif, M., Rashid, A. and Edwin, C. A. (2019). Profiling players using real-world datasets: Clustering the data and correlating the results with the big-five personality traits, *IEEE Transactions on Affective Computing* **10**(4): 568–584. https://doi.org/10.1109/TAFFC.2017.2751602.

Hamerly, G. and Elkan, C. (2004). Learning the k in k-means, *Advances in neural information processing systems* **16**: 281–288. https://dl.acm.org/doi/10.5555/2981345.2981381.

Hand, D. J. (2006). *Data Mining*, Vol. 2, Wiley Online Library.

Hand, D. J. (2007). Principles of data mining, *Drug safety* **30**(7): 621–622. https://doi.org/10.2165/00002018-200730070-00010.

Isaak, J. and Hanna, M. J. (2018). User data privacy: Facebook, cambridge analytica, and privacy protection, *Computer* **51**(8): 56–59. https://doi.org/10.1109/MC.2018.3191268.

Ismaili, O. A., Lemaire, V. and Cornuéjols, A. (2014). A supervised methodology to measure the variables contribution to a clustering, *International Conference on Neural Information Processing*, Springer, pp. 159–166. https://doi.org/10.1007/978-3-319-12637-1_20.

Jachimowicz, J., Matz, S. and Polonski, V. (2017). The behavioral scientist's ethics checklist, *Behavioral Scientist* . Disponível em https://behavioralscientist.org/behavioral-scientists-ethics-checklist/.

Jawlik, A. A. (2016). *Statistics from a to z: Confusing concepts clarified*, John Wiley & Sons.

Kaushal, V. and Patwardhan, M. (2018). Emerging trends in personality identification using online social networks—a literature survey, *ACM Transactions on Knowledge Discovery from Data (TKDD)* **12**(2): 1–30. https://doi.org/10.1145/3070645.

Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D. and Graepel, T. (2014). Manifestations of user personality in website choice and behaviour on online social networks, *Machine learning* **95**(3): 357–380. https://doi.org/10.1007/s10994-013-5415-y.

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V. and Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines, *American Psychologist* **70**(6): 543–556. https://doi.org/10.1037/a0039210.

Kosinski, M., Stillwell, D. and Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior, *Proceedings of the national academy of sciences* **110**(15): 5802–5805. https://doi.org/10.1073/pnas.1218772110.

Kuss, D. J. and Griffiths, M. D. (2011). Online social networking and addiction—a review of the psychological literature, *International journal of environmental research and public health* **8**(9): 3528–3552. https://doi.org/10.3390%2Fijerph8093528.

Lambiotte, R. and Kosinski, M. (2014). Tracking the digital footprints of personality, *Proceedings of the IEEE* **102**(12): 1934–1939. https://doi.org/10.1109/JPROC.2014.2359054.

Laros, J. A., Peres, A. J. d. S., Andrade, J. M. d. and Passos, M. F. D. (2018). Validity evidence of two short scales measuring the big five personality factors, *Psicologia: Reflexão e Crítica* **31**. https://doi.org/10.1186/s41155-018-0111-2.

Li, Y. and Wu, H. (2012). A clustering method based on k-means algorithm, *Physics Procedia* **25**: 1104–1109. https://doi.org/10.1016/j.phpro.2012.03.206.

Maldonado, V. N., Blum, R. O. and Borelli, A. (2019). *LGPD: Lei geral de proteção de dados: comentada*, Revista dos Tribunais.

Marengo, D. and Settanni, M. (2019). Mining facebook data for personality prediction: An overview, *Digital Phenotyping and Mobile Sensing*, Springer, pp. 109–124. https://doi.org/10.1007/978-3-030-31620-4_7.

Markovikj, D., Gievska, S., Kosinski, M. and Stillwell, D. (2021). Mining facebook data for predictive personality modeling, Vol. 7, pp. 23–26. Disponível em https://ojs.aaai.org/index.php/ICWSM/article/view/14466.

Matz, S. C., Appel, R. E. and Kosinski, M. (2020). Privacy in the age of psychological targeting, *Current opinion in psychology* **31**: 116–121. https://doi.org/10.1016/j.copsyc.2019.08.010.

Matz, S. C., Kosinski, M., Nave, G. and Stillwell, D. J. (2017). Psychological targeting as an effective approach to digital mass persuasion, *Proceedings of the national academy of sciences* **114**(48): 12714–12719. https://doi.org/10.1073/pnas.1710966114.

Matz, S. and Kosinski, M. (2019). Using consumers' digital footprints for more persuasive mass communication, *NIM Marketing Intelligence Review* **11**(2): 18–23. Disponível em https://www.nim.org/en/dokument/2019miraiarticle-dgital-footprinteng.

McCrae, R. R. and Costa, P. T. (2003). *Personality in adulthood: A five-factor theory perspective*, Guilford Press.

Muhammad, S. S., Dey, B. L. and Weerakkody, V. (2018). Analysis of factors that influence customers' willingness to leave big data digital footprints on social media: A systematic review of literature, *Information Systems Frontiers* **20**(3): 559–576. https://doi.org/10.1007/s10796-017-9802-y.

Ortigosa, A., Carro, R. M. and Quiroga, J. I. (2014). Predicting user personality by mining social interactions in facebook, *Journal of computer and System Sciences* **80**(1): 57–71. https://doi.org/10.1016/j.jcss.2013.03.008.

Santini, M. (2016). Advantages & disadvantages of k-means and hierarchical clustering (unsupervised learning), *Department of Linguistics and Philology, Uppsala University* . Disponível em http://santini.se/teaching/ml/2016/Lect_10/10c_UnsupervisedMethods.pdf.

Scherer, K. R. (2005). What are emotions? and how can they be measured?, *Social science information* **44**(4): 695–729. https://doi.org/10.1177%2F0539018405058216.

Segalin, C., Celli, F., Polonio, L., Kosinski, M., Stillwell, D., Sebe, N., Cristani, M. and Lepri, B. (2017). What your facebook profile picture reveals about your personality, *Proceedings of the 25th ACM international conference on Multimedia*, pp. 460–468. https://doi.org/10.1145/3123266.3123331.

Simon, H. A. (1967). Motivational and emotional controls of cognition, *Psychological review* **74**(1): 29–39. https://psycnet.apa.org/doi/10.1037/h0024127.

Smith, A. (2018). Public attitudes toward computer algorithms, *Pew Research Center* . Disponível em https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms/.

Statista (2020). Number of social media users worldwide from 2010 to 2021 (in billions). Disponível em https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/.

Stillwell, D. J. and Kosinski, M. (2004). mypersonality project: Example of successful utilization of online social networks for large-scale social research, *American Psychologist* **59**(2): 93–104. Disponível em http://www.davidstillwell.co.uk/articles/Stillwell_and_Kosinski_(2012)_myPersonality_Introduction.pdf.

Tikkinen-Piri, C., Rohunen, A. and Markkula, J. (2018). EU general data protection regulation: Changes and implications for personal data collecting companies, *Computer Law & Security Review* **34**(1): 134–153. https://doi.org/10.1016/j.clsr.2017.05.015.

Verma, D. and Meila, M. (2003). A comparison of spectral clustering algorithms, *University of Washington Tech Rep UWCSE030501* **1**: 1–18. Disponível em https://sites.stat.washington.edu/spectral/papers/UW-CSE-03-05-01.pdf.

Vinciarelli, A. and Mohammadi, G. (2014). A survey of personality computing, *IEEE Transactions on Affective Computing* **5**(3): 273–291. https://doi.org/10.1109/TAFFC.2014.2330816.

Wald, R., Khoshgoftaar, T. and Sumner, C. (2012). Machine prediction of personality from facebook profiles, *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, IEEE, pp. 109–115. https://doi.org/10.1109/IRI.2012.6302998.

Walfish, S. (2006). A review of statistical outlier methods, *Pharmaceutical technology* **30**(11): 82. Disponível em http://www.statisticaloutsourcingservices.com/Outlier2.pdf.

Williams, L. and Pennington, D. (2018). An authentic self: Big data and passive digital footprints, *International Symposium on Human Aspects of Information Security & Assurance (HAISA 2018)*. Disponível em https://strathprints.strath.ac.uk/65205/.

Youyou, W., Kosinski, M. and Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans, *Proceedings of the National Academy of Sciences* **112**(4): 1036–1040. https://doi.org/10.1073/pnas.1418680112.

Zeng, Z., Hu, Y., Roisman, G. I., Wen, Z., Fu, Y. and Huang, T. S. (2007). Audio-visual spontaneous emotion recognition, *Artifical Intelligence for Human Computing*, Springer, pp. 72–90. https://doi.org/10.1007/978-3-540-72348-6_4.