The Future of Data-driven Personas: A Marriage of Online Analytics Numbers and Human Attributes

Joni Salminen[1,2], Soon-gyo Jung[1] and Bernard J. Jansen[1]
[1]*Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar*
[2]*Turku School of Economics at the University of Turku, Turku, Finland*

Keywords: Data-driven Personas, Automatic Persona Generation, Online Analytics, Customer Segmentation, Marketing, Big Data, Automation.

Abstract: The massive volume of online analytics data about customers has led to novel opportunities for user segmentation. However, getting real value from data remains challenging for many organizations. One of the recent innovations in online analytics is data-driven persona generation that can be used to create high-quality human representations from online analytics data. This manuscript (a) summarizes the potential of data-driven persona generation for online analytics, (b) characterizes nine open research questions for data-driven persona generation, and (c) outlines a research agenda for making persona analytics more useful for decision makers.

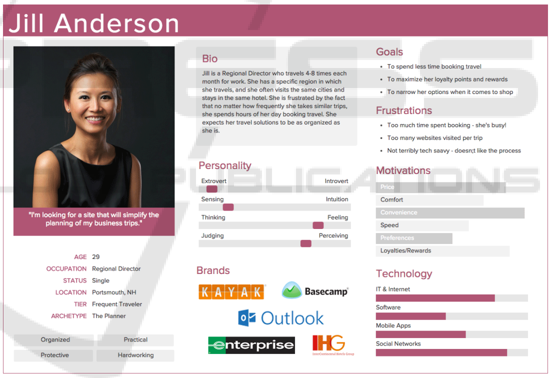
1 INTRODUCTION

Despite the increasing availability of online analytics data (also referred to as “Big Data”), decision makers are trying to turn customer data into practical insights (Salminen et al., 2017a). For this reason, various approaches for automatic analytics and insight generation have been proposed (Salminen and Jansen, 2018; Wang et al., 2018).

One approach for better understanding customers is the persona technique, popularized by Cooper (2004). A persona is defined as a *fictitious person representing an underlying customer or user group*, often the core customers of an organization, although they can also be the potential or desired users of a system (Cooper, 2004) (see Figure 1 for an example). Personas are deployed for various purposes, e.g., software development, design, marketing, and health informatics (Goodwin and Cooper, 2009). They facilitate the communication of data within an organization, so that decisions can be made keeping the customers in mind (Long, 2009).

From the analytics perspective, personas segment similar customers under one *archetype*, aiding decision makers to understand customer needs and wants. While it is not practical to cognitively process thousands of individuals for customer decisions, a few core customer segments is feasible for humans.

As online analytics data has become more prevalent and accessible, researchers have proposed novel methods for *data-driven persona generation* that uses digital, rather than analog, data for persona creation (Zhang et al., 2016).



Jill Anderson

"I'm looking for a site that will simplify the planning of my business trips."

AGE	29
OCCUPATION	Regional Director
STATUS	Single
LOCATION	Portsmouth, NH
TIER	Frequent Traveler
ARCHETYPE	The Planner

Organized Practical
Protective Hardworking

Bio

Jill is a Regional Director who travels 4-8 times each month for work. She has a specific region in which she travels, and she often visits the same cities and stays in the same hotel. She is frustrated by the fact that no matter how frequently she takes similar trips, she spends hours of her day booking travel. She expects her travel solutions to be as organized as she is.

Personality

Brands

Goals

- To spend less time booking travel
- To maximize her loyalty points and rewards
- To narrow her options when it comes to shop

Frustrations

- Too much time spent booking - she's busy!
- Too many websites visited per trip
- Not terribly tech saavy - doesn;t like the process

Motivations

Technology

Figure 1: Example of a persona profile[1]. A typical persona profile has a name, picture, and text description of the persona.

Data-driven persona generation addresses two major challenges in persona creation: (a) *the complexity and cumbersomeness of using large amounts of customer data for creating personas*, and (b) *the slow and expensive process of creating personas manually*. Data-driven personas transform online analytics data into representations that decision makers can easily process (An et al., 2018c; Salminen et al., 2018a). Further, data-driven personas are created rapidly and updated easily, while preserving the privacy of individuals (An et al., 2018b).

Salminen, J., Jung, S. and Jansen, B.
The Future of Data-driven Personas: A Marriage of Online Analytics Numbers and Human Attributes.
DOI: 10.5220/0007744706080615
In *Proceedings of the 21st International Conference on Enterprise Information Systems (ICEIS 2019)*, pages 608-615
ISBN: 978-989-758-372-8

However, to generate useful and accurate data-driven personas *without any manual interventions*, there are several challenges to address. These challenges relate to sub-fields of computer science, such as Image Generation, Natural Language Processing, Topic Modeling, Algorithms, and Human-Computer Interaction, as well as various "softer" topics such as persona perceptions (Salminen et al., 2018c), persona biases (Hill et al., 2017; Salminen et al., 2019b), and value in use (Salminen et al., 2018b).

In this manuscript, we explore a contemporary collection of research challenges related to data-driven persona creation, particularly from the perspective of automatic persona generation. We aim to inspire research within persona studies and related subfields of computer science.

2 THE RISE OF DATA-DRIVEN PERSONAS

2.1 Limitations of Manual Persona Generation

Personas are typically created with qualitative approaches. Brickey et al. (2012) found that 81% of persona creation efforts reported in academic literature have applied qualitative techniques, such as ethnographic fieldwork and interviews. However, manual persona generation has been thoroughly criticized in the literature, the main criticism being:

Non-Representative Data: Manually created personas typically rely on data that does not represent the whole customer base (Chapman and Milham, 2006).

Lack of Scaling: Because manual analysis relies on human labor, it scales poorly with the big datasets used in online analytics (An et al., 2018b).

High Cost: Manual persona generation is costly; it typically takes several months and costs tens of thousands of dollars. The high cost factor keeps personas from the reach of small to medium-size businesses and start-ups.

Expiration: Personas tend to expire when changes in customer behavior take place. This is typical for many fast-moving online businesses, including online purchase behavior (Salminen et al., 2017b), search behavior (Jansen et al., 2011), and online content consumption (Abbar et al., 2015).

2.2 Advances in Data-driven Persona Generation

To solve the challenges of manual persona generation, researchers have suggested quantitative persona generation. The main techniques are as follows.

Quantitative Analysis of Survey Data: Several prior attempts for data-driven persona creation rely on survey-based data collection (Chapman et al., 2015; Dupree et al., 2016; Vahlo et al., 2017). This survey data is then most typically analyzed via cluster or factor analysis. However, survey-based data collection can be costly and fallible compared to using behavioral data due to many possible respondent and researcher biases associated to survey data collection in general (Podsakoff et al., 2003).

System Log Data: In addition to survey data, personas can be created from system logs, and organizational records describing the users or customers (Brickey et al., 2012). For example, Molenaar (2017) analyzed 400,000 clickstreams from a period of three months, grouping them into common workflows and classifying users into these workflows. Using a similar approach, Zhang et al. (2016) applied hierarchical clustering to generate five data-driven personas from clickstream data.

Procedural Personas: In video game context, researchers have created procedural personas that capture the sequential game-playing choices. The applied techniques include, e.g., evolutionary algorithms and neural networks (Holmgard et al., 2018). The procedural personas are given names based on their behaviors of playing the game (e.g., "Monster Killers"). Rather than being "rounded personas" (Nielsen, 2004) with name and demographic information, these personas can be seen as virtual agents that model the possible game-playing behaviors (Vahlo et al., 2017).

Latent Semantic Analysis (LSA): LSA has been applied to create personas by differentiating users based on their use of language (Miaskiewicz et al., 2009). The weakness of this approach is the dependency on the text corpus which is not always available in online analytics data. In addition, not using behavioral data (e.g., product engagement) can be considered as a weakness.

Discrete Choice Analysis: In the discrete choice methodology for persona creation (Chapman et al., 2015), customers explicitly state their preferences and a conjoint analysis algorithm is then used to match respondent to their best-fit persona. The method was developed to answer the criticism of personas as lacking quantitative information (Chapman et al., 2008), as it enables, through forced assignment, to determine the proportional representativeness of personas within the overall user base. The method also makes it possible to compare the algorithmic persona assignments to randomly generated persona assignments. However, the major limitation is that stated preference data can be expensive to collect and can also be more unreliable than observed behavioral data. This is also the limitation of creating personas with principal component analysis (Sinha, 2003) that uses preference data from a limited number of customers.

Automatic Persona Generation (APG): APG is defined both as *a methodology and a system for automatic creation of personas from online analytics data* (An et al., 2018b). Automatically generated personas are (1) representative, as APG processes the entire online analytics dataset, (2) behaviorally accurate, inferring patterns from customers' engagement with products (e.g., digital content, e-commerce products, flight destinations...), (3) rapidly generated due to fast processing time, and (4) constantly up-to-date due to period refreshing of the data and the associated regeneration of the personas (An et al., 2018b,c).

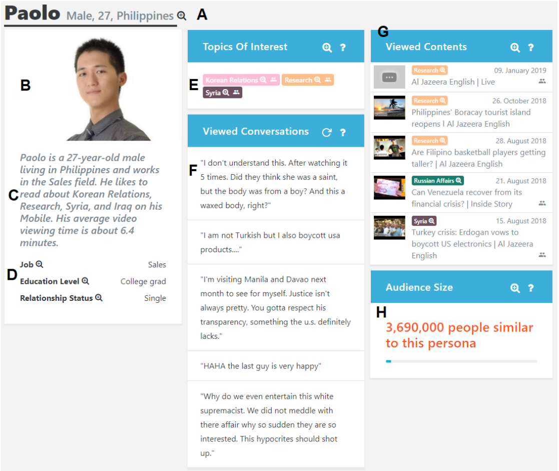
The following section explains the APG methodology. We focus on this approach as it represents the latest techniques for data-driven persona generation and specifically utilizes online analytics data.

3 AUTOMATIC PERSONA GENERATION

3.1 Data Collection from Online Analytics Platforms

Online analytics platforms (e.g., Google Analytics, YouTube Analytics, Facebook Insights) typically enable collection of user data automatically via application programming interfaces (APIs) [2]. Typically, this data is aggregated into segments to protect the privacy of individual users. An example of aggregated user segment is [Male, 44-55, Qatar]. The segments given by the online analytics platforms typically contain information of the gender, age, and country of the users. The platforms typically collect this information from the users upon registration. Various interaction metrics can be retrieved for each group (e.g., clicks, views). For example, [Female, 24-35, USA] $\longrightarrow 1,590$ views for Video A.

Using the APIs of online analytics platform, APG collects the aggregated data for products and engagement metrics. For example, from YouTube Analytics, APG collects videos and their view counts, whereas, from Google Analytics, APG collects pages and number of sessions.



Paolo Male, 27, Philippines

Paolo is a 27-year-old male living in Philippines and works in the Sales field. He likes to read about Korean Relations, Research, Syria, and Iraq on his Mobile. His average video viewing time is about 6.4 minutes.

Job	Sales
Education Level	College grad
Relationship Status	Single

Topics Of Interest ?

Korean Relations Research Syria

Viewed Conversations ?

"I don't understand this. After watching it 5 times. Did they think she was a saint, but the body was from a boy? And this a waxed body, right?"

"I am not Turkish but I also boycott usa products...."

"I'm visiting Manila and Davao next month to see for myself. Justice isn't always pretty. You gotta respect his transparency, something the u.s. definitely lacks."

"HAHA the last guy is very happy"

"Why do we even entertain this white supremacist. We did not meddle with there affair why so sudden they are so interested. This hypocrites should shot up."

Viewed Contents ?

Research 09. January 2019 Al Jazeera English | Live

Research 26. October 2018 Philippines' Boracay tourist island reopens I Al Jazeera English

Research 28. August 2018 Are Filipino basketball players getting taller? | Al Jazeera English

Russian Affairs 21. August 2018 Can Venezuela recover from its financial crisis? | Inside Story

Syria 15. August 2018 Turkey crisis: Erdogan vows to boycott US electronics | Al Jazeera English

Audience Size ?

3,690,000 people similar to this persona

Figure 2: Output of APG. Denoted areas are: [A] Name and demographic information, [B] Picture, [C] Text description, [D] Life situation information, [E] Topics of interest, [F] Social media comments, [G] Most interested content, and [F] Audience size.

3.2 Data Processing and Persona Profile Generation

After collecting the data from an online analytics platform, APG transforms it into an *interaction matrix* that captures the interaction between customers and online products (An et al., 2018c,b).

$\mathbf{V}$ denotes the $\mathbf{g} \times \mathbf{c}$ matrix of $\mathbf{g}$ customer groups and $\mathbf{c}$ online products. The element $\mathbf{V}$, V_{ij}, can be any metric that reflects the engagement of the customer group G_i for product C_j. For example, in YouTube Analytics, V_{ij} is a view count for a particular video, C_j from customer group G_i. The customer groups contain gender, age, and country (e.g., Female, 54-65, South Korea). Using $\mathbf{V}$ as the basis, non-negative matrix factorization is applied to detect $\mathbf{p}$ latent patterns (An et al., 2018c).

These patterns form the core of the personas, as they represent the customer groups' product preferences. APG then chooses a representative demographic group for each latent pattern and enriches this demographic group with additional information to produce a complete persona profile (see Figure 2).

[2] Note that accessing the analytics data requires authorization from the owner of the analytics property.

4 FUTURE OF AUTOMATIC PERSONA GENERATION

The vision of APG can be summarized as *achieving completely automatic generation of high-quality personas that addresses the limitations of manual persona creation through the employment of online analytics data.* "Completely automatic" refers to elimination of manual steps. "High quality" refers to serving the persona user's decision needs while accurately representing the underlying data about the customers.

In the subsections, we present Proposed Research Questions (PRQs) toward the vision.

4.1 Automatic Generation of Persona Pictures

In the current version of APG, a photo for each persona is purchased through online stock photo banks. However, it is difficult and costly to find an appropriate photo for all demographic groups, especially worldwide. The potential solution could be to generate persona pictures automatically.

The nascent developments in Deep Learning, particularly in Generative Adversarial Networks (GAN) Goodfellow et al. (2014), have been applied in generating and modifying human faces. This line of work could be used for generating photos that vary by persona's age, gender, and ethnicity – possibly using these attributes as conditions in the Conditional GAN architecture Isola et al. (2017). **PRQ1: How to automatically generate persona profile pictures?**

4.2 Defining the Optimal Persona Attributes

Persona profiles typically contain name, age, and gender, as well as other demographic information, such as marital status, education level, job, and so on. However, there have been few studies into what information should be included in a persona. This lack of prior work speaks to a need for user studies, including interviews and ethnographic investigations of actual users of personas in the workplace.

Related to this issue of defining the optimal information content of a persona profile, another major limitation of data-driven persona methodologies is that none of them currently infer in-depth insights about the users such as needs and wants that are essential for the thorough understanding of the users or customers that the persona portrays (Cooper, 2004).

We summarize these issues in two research questions: **PRQ2: What information should automatically generated personas contain? PRQ3: How could that information be automatically inferred from online analytics data?**

Two approaches could potentially address PRQ2: (1) defining shared information needs for a given industry, and constructing industry-specific templates (e.g., e-commerce, online media and news, e-health, etc.); or (2) providing a self-selection options for users to build personas by choosing from all available information. In the former case, the selection of persona attributes should depend on persona users' information needs that can be obtained via user studies (Salminen et al., 2018d). Overall, determining the persona users' information needs relies on implicit or explicit user feedback.

A potential solution for PRQ3 is the use of computational methods for inferring customer attributes from social media.There is a substantial amount of research using social media platforms, such as Twitter, to infer user attributes (Volkova et al., 2015). These studies have inferred, for example, social media users' psychological traits, socioeconomic status, relationship status, political orientation, and brand likings, by using profile information, comments, and connections of the user. The applied techniques are diverse, including natural language processing, graph analysis, and various machine learning classifiers. If any of the above attributes are considered critical by persona users in a specific application domain, methods of inferring those attributes and associating them with the automatically generated personas (most likely using probabilistic matching) are called for.

Moreover, these additional customer attributes could be available on-demand, so that the persona users could construct their own personas from ground-up by choosing the information elements that matter in their respective industries or use cases.

4.3 Unsupervised Learning of Persona's Interest

In addition to demographic information, online analytics data contains information on customer preferences that can be inferred from the customers' engagement with various online products. However, due to vast number of individual products, they need to be categorized in order to provide a meaningful summary of the customer preferences. Thus, data-driven personas should incorporate unsupervised topic models that can accurately classify the online products based on their features, such as textual descriptions. This prompts the research question: **PRQ4: How to generate a universal topic taxonomy for online content?** Here, unsupervised learning methods, such as Latent Dirichlet Allocation (LDA) Topic Mod-

Table 1: Research and development goals for automatic persona generation.

Persona section addressed	Proposed solution	Ideal outcome	Applicable domains
Persona Picture	Automatically producing persona face pictures for the matching demographic variables (age, gender, country)	Eliminates the need for manually acquisition of persona pictures.	Computer Vision, Generative Adversarial Networks
Topics of interest	Automatically creating a taxonomy that is scalable across multiple topical domains.	Eliminates the need for creating an organization- or industry-specific taxonomy each time a new one is added.	Topic Modeling: LDA, LSA; Entity Resolution, Data Mapping
Persona quotes	Providing comments that relevant for persona user's use case and do not distract the persona user from the important attributes of the persona.	Increases empathy and customer insights gained by the persona user; eliminates distraction caused by non-useful comments.	Social Computing; Text Classification, Natural Language Processing
Persona attributes	Determining the persona attributes that correspond to the persona users' needs in a given decision-making situation and devising methods to infer those attributes from unstructured data such as social media comments	Satisfying the persona users' information needs, thereby enabling possible better decision-making via the use of personas.	Human-Computer Interaction, Information Design, Usability
Overall persona profile	Validating accuracy, consistency, and usefulness of personas for individuals and organizations.	Ensuring that personas are reliable and valid, so that they can be trusted in real decision-making situations.	Case Studies, User Studies, HCI

elling (Hong et al., 2018), could be helpful. In addition, Google's Universal Sentence Encoder uses a hybrid approach that outputs similarity with a known taxonomy for any text content (Cer et al., 2018).

Another challenge related to inferring additional customer attributes is their association with the persona profiles generated using *different* source data. For example, *Platform A* has information on a persona's topics of interest, and *Platform B* has information on the persona's movie preferences. Then, there is a need for mapping these seemingly disconnected pieces of information in order to include both of them in the persona profile. To create high-quality personas with attributes such as the persona's goals, needs, and wants, several data sources need to be combined. Therefore, **PRQ5: How can we map the personas to online users across different platforms?** Approaches studied in the domain of entity resolution could be of help here.

4.4 Choosing High-quality Quotes for the Persona

Descriptive quotes are typically shown in the persona profile to provide a better understanding of the customers (Cooper, 2004). However, it has been found that the quotes can also distract the persona users toward information that is not relevant for their task. For example, Salminen et al. (2018d) found that the ethnicity of the persona affected the persona users' interpretation of the persona. Moreover, when showing unfiltered social media comments as persona quotes, the impression of the persona can quickly turn toxic (Salminen et al., 2018d). To counter this issue, Salminen et al. (2019a) have proposed three criteria for the automatic selection social media comments as persona quotes:

1. *Representativeness:* the selected comments match the behavioral patterns, topics of interest, and demographics of the corresponding persona
2. *Relevance:* the selected comments are helpful for the persona user in his or her purpose for using the persona
3. *Non-toxicity:* the selected comments are not offensive to the degree where they would distract the persona user from the other information in the persona profile.

The associated solutions require filtering out the most toxic comments by automatic classification (or use of dictionary-based methods), but also mapping the comments with the matching personas. In a recent workshop on automatic persona generation, it was suggested that the mapping could be done based on demographic analysis of the social media users' profile information (when publicly available) or by inferring their gender, age and interest from the style of their writing (An et al., 2018a). Thus, the research questions are: **PRQ6: How can the the attributes of the commenting customers be inferred only using the text of the social media comments (when no public profile information is available) to select the comments that meet a persona's attributes? PRQ7: How to select the most relevant comments to the end user?**

Moreover, in filtering out toxic comments, we should be cautious of manipulating the actual data and thereby biasing the information shown in the persona profile. Therefore, if the data in fact contains a high number of toxic comments, to maintain the truthfulness of the persona, data-driven personas should display those comments, even if some end users might find them offensive. Thus, the challenge of toxicity in personas involves a certain trade-off between truthfulness and user experience.

4.5 Avoiding Biasing the End Users of Personas

One challenge of the personas is the fact that for the selected attributes, only one dominant value can be chosen. For example, the persona can have only one age, even though the customers that the particular persona represents form a distribution of ages. This concern is highlighted in data-driven persona creation methodologies that are based on behavioral or preference patterns, because many demographic groups can share behavioral patterns or preferences. Choosing one dominant value for an attribute, say, gender or ethnicity, can easily result in biased interpretations by persona users (Salminen et al., 2018d). Thus, **PRQ8: How can personas be debiased so that oversimplication of the customer base is avoided?**

Two solutions can be thought of: (1) removing ambiguous informational to debias the persona for end users, and (2) purposefully introducing diversity to display the variation in the underlying user base. For example, it is possible to introduce an additional layer of information in the persona profiles (Salminen et al., 2019a). Such an approach could be used to mirror each active information element in a "deeper layer" that holds breakdown information. By showing deeper information, it may be possible to curb the tendency of personas to evoke stereotypical thinking.

The drawback of this approach is that it may reduce the empathy-benefits of persona (immersion, understanding) (Cooper, 2004), so that instead of being a believable person, the persona becomes a fragmented group of different people. To maintain the credibility of the persona, a coherence of the whole is needed. These perceptual questions are conceptually linked to evaluation of the persona profiles, an area that is critical for adoption and real use of personas in organizations. Toward that end, our final research question is **PRQ9: How to evaluate the usefulness and value generated by data-driven personas?**

5 EVALUATING THE QUALITY OF DATA-DRIVEN PERSONAS

Finally, it is not immediately evident how to measure the quality of data-driven personas. For example, how can their accuracy (in terms of correspondence with the data) be verified? Is accuracy even correlated with the perceived usefulness of the personas? In disentangling these questions, researchers have mostly focused on the technical aspects of persona quality (Chapman and Milham, 2006; Chapman et al., 2008). Yet, there is a nascent stream of studies focusing on persona perceptions (Marsden and Haag, 2016; Hill et al., 2017; Salminen et al., 2018d).

For example, Salminen et al. (2018c) developed a Persona Perception Scale that lists several perceptual constructs associated with the use of personas. From this scale, at least the following ones could be perceived important for evaluating persona quality: credibility, consistency, completeness, and clarity. In order for personas to be useful, the persona users need to perceive them as credible (i.e., trustworthy, reliable). Moreover, the information in the data-driven persona profiles needs to be consistent (e.g., topics of interest need to match the quotes), or else there is a risk of confusion among the persona users. In turn, if the personas are not complete (i.e., contain all the necessary information that the persona user needs for accomplishing their task), they can hardly be considered useful. Finally, information should be presented clearly; for example, unclear titles or description for the persona content sections are likely to cause confusion among end users (Salminen et al., 2018d).

6 DISCUSSION

Data-driven personas of the future should be low-cost, accurate, and accessible by small and large organizations with varying budgets and needs. However, many challenges await before reaching this vision.

To investigate these challenges, we formulated nine research questions that deal with various aspects of automatic persona generation. Addressing these questions, we believe, would result in major progress toward creating high-quality personas from customer analytics data. This goal is impactful for real organizations deploying personas for use cases such as product development, design, and marketing.

This manuscript represents a call for action to researchers interested in "humanizing" online analytics, encouraging contributions in methodological and practical development of data-driven personas. We expect that addressing the research questions proposed here requires several years of active research, with the potential of several new avenues of inquiry in multiple domains. Data-driven persona creation is an on-going research field with potential for both focused disciplinary and cross-disciplinary research in Algorithms, HCI, Online Analytics, and so on.

Personas are also opening new research avenues for experiments in Computational Social Science, particularly revealing end users' subjective perceptions and biases about the audience or user groups (Hill et al., 2017; Salminen et al., 2018d). By classifying personas according to their attributes (e.g., age, gender, ethnicity), it is possible conduct user studies that examine how end users perceive and respond to different personas. Another line of research is to investigate the possibility of algorithmic bias in the automatically generated personas.

While automatically generated personas may not replace numbers in online analytics, they do provide intuitive descriptions of the customer base using quantitative data. In the APG system (Jung et al., 2017), numbers remain available as raw data that can be downloaded by the end users and as data breakdowns. Thus, data-driven personas can support decision making by providing humanlike renderings of numerical customer data, while providing an access to the underlying raw data.

7 CONCLUSION

The advancements in machine learning and Web technologies, combined with online analytics data, show great promise for data-driven persona generation. With these novel methods, it becomes possible to bring personas in the reach of more decision makers in more organizations, enhancing customer-oriented decision making and democratizing personas for all organizations, including corporations, small businesses, and startups using online analytics data.

REFERENCES

Abbar, S., An, J., Kwak, H., Messaoui, Y., and Borge-Holthoefer, J. (2015). Consumers and suppliers: Attention asymmetries. a case study of al jazeera's news coverage and comments. In *Computational Journalism Symposium*.

An, J., Chunara, R., Crandall, D. J., Frajberg, D., French, M., Jansen, B. J., Kulshrestha, J., Mejova, Y., Romero, D. M., Salminen, J., Sharma, A., Sheth, A., Tan, C., Taylor, S. H., and Wijeratne, S. (2018a). Reports of the Workshops Held at the 2018 International AAAI Conference on Web and Social Media. *AI Magazine*, 39(4):36–44.

An, J., Kwak, H., Jung, S.-g., Salminen, J., and Jansen, B. J. (2018b). Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, 8(1).

An, J., Kwak, H., Salminen, J., Jung, S.-g., and Jansen, B. J. (2018c). Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Transactions on the Web (TWEB)*, 12(3).

Brickey, J., Walczak, S., and Burgess, T. (2012). Comparing semi-automated clustering methods for persona development. *IEEE Transactions on Software Engineering*, 38(3):537–546.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., and Tar, C. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

Chapman, C., Krontiris, K., and Webb, J. (2015). Profile CBC: Using Conjoint Analysis for Consumer Profiles. Technical report, Google Research.

Chapman, C., Love, E., Milham, R., ElRif, P., and Alford, J. (2008). Quantitative evaluation of personas as information. In *Human Factors and Ergonomics Society 52nd Annual Meeting*, pages 1107–1111.

Chapman, C. N. and Milham, R. P. (2006). The personas' new clothes: Methodological and practical arguments against a popular method. In *Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 634–636.

Cooper, A. (2004). *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2nd Edition)*. Pearson Higher Education.

Dupree, J. L., Devries, R., Berry, D. M., and Lank, E. (2016). Privacy personas: Clustering users via attitudes and behaviors toward security practices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI2016)*, pages 5228–5239. ACM.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

Goodwin, K. and Cooper, A. (2009). *Designing for the Digital Age: How to Create Human-Centered Products and Services*. Wiley, Indianapolis, IN.

Hill, C. G., Haag, M., Oleson, A., Mendez, C., Marsden, N., Sarma, A., and Burnett, M. (2017). Gender-inclusiveness personas vs. stereotyping: Can we have it both ways? In *Proceedings of the ACM Conference* *on Human Factors in Computing Systems (CHI2017)*, pages 6658–6671. ACM Press.

Holmgard, C., Green, M. C., Liapis, A., and Togelius, J. (2018). Automated Playtesting with Procedural Personas with Evolved Heuristics. *IEEE Transactions on Games*, PP(99):1–1.

Hong, W., Zheng, X., Qi, J., Wang, W., and Weng, Y. (2018). Project Rank: An Internet Topic Evaluation Model Based on Latent Dirichlet Allocation. In *2018 13th International Conference on Computer Science & Education (ICCSE)*, pages 1–4. IEEE.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE.

Jansen, B. J., Sobel, K., and Cook, G. (2011). Classifying ecommerce information sharing behaviour by youths on social networking sites. *Journal of Information Science*, 37(2):120–136.

Jung, S.-G., An, J., Kwak, H., Ahmad, M., Nielsen, L., and Jansen, B. J. (2017). Persona generation from aggregated social media data. In *Extended Abstracts on Human Factors in Computing Systems (CHI2017)*, pages 1748–1755.

Long, F. (2009). Real or imaginary: The effectiveness of using personas in product design. In *Proceedings of the Irish Ergonomics Society Annual Conference*, volume 14. Irish Ergonomics Society Dublin.

Marsden, N. and Haag, M. (2016). Stereotypes and politics: Reflections on personas. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI2016)*, pages 4017–4031. ACM.

Miaskiewicz, T., Grant, S. J., and Kozar, K. A. (2009). A preliminary examination of using personas to enhance user-centered design. In *AMCIS 2009 Proceedings*, page Article 697.

Molenaar, L. (2017). Data-driven personas: Generating consumer insights with the use of clustering analysis from big data. Master's thesis, TU Delft, Netherlands.

Nielsen, L. (2004). *Engaging Personas and Narrative Scenarios*. Doctoral dissertation. Copenhagen Business School.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., and Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5):879.

Salminen, J., Şengün, S., Kwak, H., Jansen, B. J., An, J., Jung, S.-g., Vieweg, S., and Harrell, F. (2018a). From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas. *First Monday*, 23(6).

Salminen, J. and Jansen, B. J. (2018). Use Cases and Outlooks for Automatic Analytics. *arXiv:1810.00358 [cs]*. arXiv: 1810.00358.

Salminen, J., Jansen, B. J., An, J., Kwak, H., and Jung, S.-g. (2018b). Are personas done? Evaluating their usefulness in the age of digital analytics. *Persona Studies*, 4(2):47–65.

Salminen, J., Jansen, B. J., An, J., Kwak, H., and Jung, S.-G. (2019a). Automatic Persona Generation for Online Content Creators: Conceptual Rationale and a Research Agenda. In Nielsen, L., editor, *Personas - User Focused Design*, Human–Computer Interaction Series, pages 135–160. Springer London, London.

Salminen, J., Jung, S.-G., and Jansen, B. J. (2019b). Detecting Demographic Bias in Automatically Generated Personas. In *Extended Abstracts on CHI Conference on Human Factors in Computing Systems (CHI2019)*, Glasgow, UK.

Salminen, J., Kwak, H., Santos, J. M., Jung, S.-G., An, J., and Jansen, B. J. (2018c). Persona Perception Scale: Developing and Validating an Instrument for Human-Like Representations of Data. In *Extended Abstracts on ACM Conference on Human Factors in Computing Systems (CHI2018)*, Montréal, Canada.

Salminen, J., Milenković, M., and Jansen, B. J. (2017a). Problems of data science in organizations: An explorative qualitative analysis of business professionals' concerns. In *Proceedings of International Conference on Electronic Business (ICEB 2017)*.

Salminen, J., Nielsen, L., Jung, S.-G., An, J., Kwak, H., and Jansen, B. J. (2018d). "Is More Better?": Impact of Multiple Photos on Perception of Persona Profiles. In *Proceedings of ACM Conference on Human Factors in Computing Systems (CHI2018)*, Montréal, Canada.

Salminen, J., Seitz, S., Jansen, B. J., and Salenius, T. (2017b). Gender Effect on E-Commerce Sales of Experience Gifts: Preliminary Empirical Findings. In *Proceedings of International Conference on Electronic Business (ICEB 2017)*, Dubai.

Sinha, R. (2003). Persona development for information-rich domains. In *Extended abstracts of the ACM Conference on Human Factors in Computing Systems (CHI 2003)*, pages 830–831. ACM.

Vahlo, J., Kaakinen, J. K., Holm, S. K., and Koponen, A. (2017). Digital Game Dynamics Preferences and Player Types. *Journal of Computer-Mediated Communication*, 22(2):88–103.

Volkova, S., Bachrach, Y., Armstrong, M., and Sharma, V. (2015). Inferring Latent User Properties from Texts Published in Social Media. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 4296–4297, Austin, Texas. AAAI Press.

Wang, Y., Kung, L., and Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, 126:3–13.

Zhang, X., Brown, H.-F., and Shankar, A. (2016). Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI2016)*, pages 5350–5359. ACM.