



# Finetuning Analytics Information Systems for a Better Understanding of Users: Evidence of Personification Bias on Multiple Digital Channels

Bernard J. Jansen<sup>1,2</sup> · Soon-gyo Jung<sup>1</sup> · Joni Salminen<sup>3</sup>

Accepted: 26 March 2023  
© The Author(s) 2023

## Abstract

Although the effect of hyperparameters on algorithmic outputs is well known in machine learning, the effects of hyperparameters on information systems that produce user or customer segments are relatively unexplored. This research investigates the effect of varying the number of user segments on the personification of user engagement data in a real analytics information system, employing the concept of persona. We increment the number of personas from 5 to 15 for a total of 330 personas and 33 persona generations. We then examine the effect of changing the hyperparameter on the gender, age, nationality, and combined gender-age-nationality representation of the user population. The results show that despite using the same data and algorithm, varying the number of personas strongly biases the information system's personification of the user population. The hyperparameter selection for the 990 total personas results in an average deviation of 54.5% for gender, 42.9% for age, 28.9% for nationality, and 40.5% for gender-age-nationality. A repeated analysis of two other organizations shows similar results for all attributes. The deviation occurred for all organizations on all platforms for all attributes, as high as 90.9% in some cases. The results imply that decision makers using analytics information systems should be aware of the effect of hyperparameters on the set of user or customer segments they are exposed to. Organizations looking to effectively use persona analytics systems must be wary that altering the number of personas could substantially change the results, leading to drastically different interpretations about the actual user base.

**Keywords** Personas · Hyperparameters · Analytic bias · Machine learning

## 1 Introduction

Organizations typically want to know more about their users and customers to make more informed design and business decisions. In pursuit of this goal, organizations often turn to user analytics (Fan et al., 2015; Jansen et al., 2009) and information systems that process analytics data (Shmueli &

Koppius, 2011) which we refer to as *analytics information systems* (further, to clarify our terminology, in this study we focus on *persona analytics systems*, a special case of analytics information systems that use personas as the format of presenting end-user segments to decision makers). The process of using these systems relies on user metrics that are proxies for real users for user populations that can number millions or more for large organizations. As machine learning (ML) and AI are increasingly integrated into these persona analytics systems, generating personas from large amounts of user data becomes increasingly accessible. In turn, organizations employ user analytics metrics for various purposes (Wedel & Kannan, 2016; Xu et al., 2016; Żbikowski & Antosiuk, 2021; Zhang et al., 2011), such as design, marketing, advertising, product development, upselling, and customer relationship management (CRM), and other managerial decision making. These user analytics metrics exemplify *depersonification*, which we define in this context as the representation of real people by numerical, text, or other data, creating quantitative proxies for real

---

✉ Bernard J. Jansen  
jjansen@acm.org

Soon-gyo Jung  
sjung@hbku.edu.qa

Joni Salminen  
joni.salminen@utu.fi

<sup>1</sup> Qatar Computer Research Institute, Hamad Bin Khalifa University, Doha, Qatar

<sup>2</sup> Education City, Doha, Qatar

<sup>3</sup> School of Marketing and Communication, University of Vaasa, Vaasa, Finland

users of systems, apps, products, and other offerings in the electronic marketplace.

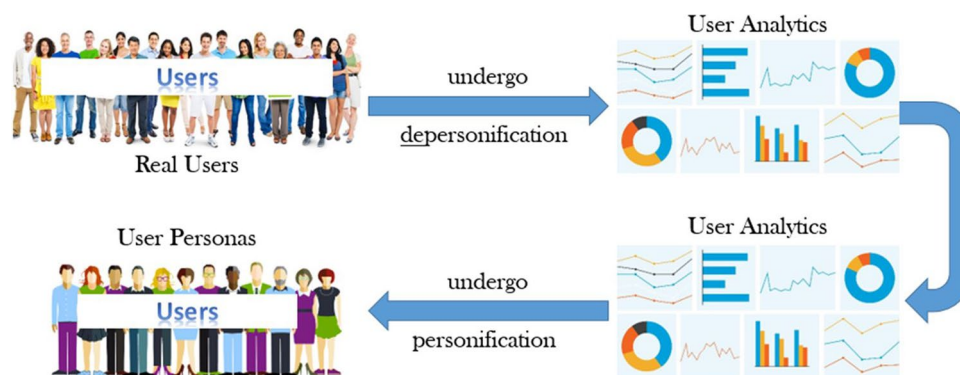
With the increasing ease of collecting data on user behaviors — such as information system interactions, online purchases, chat logs, and social media information — the data volume about users has dramatically increased to a size that benefits from the employment of ML models and data science algorithms (Agarwal & Dhar, 2014) to process and analyze user data (Lin & Kunnathur, 2019) into meaningful representations (Park & Kang, 2022). Typically, ML modeling is used with large volumes of user data to identify trends and segments within large user populations (Arora & Malik, 2015). Robust analytics information systems also use these ML models to create algorithmically-generated personas (Jung et al., 2018), which are humanized representations of people based on user data. These algorithmically-generated personas are an example of *personification* (Stevenson & Mattson, 2019), defined here as the algorithmically generated representation of numbers, text, metrics, and other user data in the form of fictitious humans.

Personification proposes interesting challenges for analytics information systems that rely on algorithmic approaches. One challenge, and the one addressed explicitly in this research, is how many algorithmically-generated personas to create during the personification process. Selecting the appropriate number of personas has parallels with the hyperparameter selection in ML, which we discuss below. Our premise is that the hyperparameter values will affect the personification process, altering the outcome of this process to a significant extent.

The motivation for the personification of user data (Delbaere et al., 2011) is that cold, rational numbers often do not generate the connection to and empathy of users required (Cohen, 2014) in endeavors such as product design, service blueprinting, marketing, advertising, and content creation – in other words, in many tasks related to

information systems in the field of analytics, i.e., analytics information systems. So, we are presented with the fascinating conundrum of organizations employing depersonification via user analytics to make more informed decisions about users. These same organizations then use personification via personas as a form of user segmentation (Wang, 2022) to better relate to and empathize with users. Figure 1 illustrates this *depersonification-personification concept*.

The *depersonification* and *personification* processes typically employ ML approaches to analyze user analytics within information systems (Griva et al., 2018) or generate algorithmically-generated personas (An et al., 2018a, b; Molenaar, 2017; Yoon et al., 2021) due to many organizations' large volumes of user data. While there are many ML approaches, including supervised and unsupervised learning, nearly all ML models require the configuration of at least two types of parameters: model parameters and hyperparameters. A *model parameter*, usually referred to as just a parameter, is a configuration variable *internal* to the model where one can estimate or calculate the value from data (e.g., mean, median, variance, maximum, minimum) with known weakness (Chen & Liu, 1993). Conversely, a *hyperparameter* is a configuration variable *external* to the model where one cannot directly extract the value from data (e.g., number of clusters, number of iterations). Hyperparameters are set manually, typically using heuristics, a priori values based on domain knowledge, or results from trial and error (e.g., using grid search techniques (Gil et al., 2019)). This setting procedure is used because one usually does not know beforehand the ideal value for a hyperparameter for a given model, dataset, and problem. Therefore, ML models (Alzubi et al., 2018) are often tuned to find the best or at least a reasonable value (Terragni & Fersini, 2021) for a model's hyperparameters that results in the best performance in context (Kalliola et al., 2021).



**Fig. 1** Illustration of the ongoing depersonification process to generate user analytics motivated by a desire to make more informed decisions about users. Parallel with this depersonification process, a personification process is ongoing where user analytics data is trans-

formed into algorithmically-generated personas motivated by a desire to make empathic and understanding decisions about users. Both processes employ similar ML approaches for organizations with large user populations

In persona analytics systems, ML models are used to create personas during the *personification* process of user modeling. Personas were traditionally created manually based on limited user data, requiring no ML approaches. However, persona analytics systems can theoretically create an arbitrary number of personas (Jansen et al., 2019b) from the user analytics of thousands to millions of users. As an emerging field of research, the impact of hyperparameters on models employed by persona analytics systems, i.e., systems that utilize user analytics information to generate personas for decision makers' user-centric tasks, has seen little research. The critical hyperparameter is the number of personas to create when deploying algorithmically-generated personas. Therefore, several unanswered questions exist in this personification domain, including: *What is the effect of changing the number of personas?*; *Does the accuracy of user representation correlate with an increase or decrease in the number of personas?*; *Is there a number of personas that is the 'best'?* These are some of the questions that motivate our research, as for a layperson using personas or even for a computational social sciences researcher with basic experience in personas and ML, the answers to these questions would not be self-evident or trivial.

These questions also matter because hyperparameter selection is a potential source of bias in algorithmic user segmentation. While there has been increased interest in algorithmic bias in various contexts (Dodge et al., 2019; Drozdal et al., 2020; Hajian et al., 2016; Zehlike et al., 2022), it has not been extensively investigated in the context of algorithmically-generated personas or user segmentation overall. One of the reasons why analyzing the impact of hyperparameters on information systems is important is to increase trust in these systems (Drozdal et al., 2020), which we particularly address in the personification and user segmentation areas.

The findings also matter for the reliability of decision-making outcomes when using user representations. Although we do not explicitly investigate this, a major justification for our research is that when aggregating user data into user representations (e.g., segments, personas), the specific composition of these representations guide decision makers' thinking and the choices they make about users. In other words, the process is as follows:

Hyperparameter selection → User representation → Decision making

For example, suppose the algorithmically-generated personas created using a persona analytics system highlight male attributes. In that case, decision makers are more likely to presume that the most important target group is males and make decisions accordingly. Of course, they could also presume the opposite: that females are under-represented and therefore should be targeted more vigorously; but in both cases, the

user representations the decision makers are exposed to frame their thinking about the users. The main point is that analytics information systems influence how users (e.g., customers) are interpreted by decision makers using these information systems for understanding the user segments. Several effects have been observed to this effect, including confirmation or validation bias (Nickerson, 1998). In our case, we define *personification bias* as the effect of analytics information on the information about the users it conveys to decision makers.

This research employs several "big" user analytics datasets numbering in the hundreds of thousands to millions of users and tens of millions of interactions across three industry standard analytics information systems (Facebook Insights, Google Analytics, Instagram Insights). We generate sets of 5 to 15 personas — which is a range in which most persona sets in the literature occur (An et al., 2018a, b) — from these user analytics datasets using an identical ML personalization approach. We then analyze the resulting 330 personas (3 systems × 110 personas per system), comparing each of the 11 groups of personas along four attributes to determine the most appropriate hyperparameter for creating personas. We then repeat the exact analysis for two separate organizations, for a total of 990 personas and 33 hyperparameter persona generations for three organizations, each on three analytics systems. From a set of personas, organizational decision makers would then focus on one or more personas to gain user understanding.

In the remainder of this work, we present a literature review, research questions, methodology, and results in the following sections. We then discuss theoretical and practical implications, and we end with a critique of this study and directions for future research.

## 2 Related Literature

### 2.1 Digital User Analytics

User analytics data is used for a variety of purposes (Kitchens et al., 2018) to support the analysis and reporting concerning the users of an organization's system, service, or product (Bijmolt et al., 2010; Griva et al., 2021; Hossain et al., 2020). Analytics about users is generally represented as numbers, even when the focus of the study is non-numeric (e.g., sentiment analysis of textual data). These numbers are presented as counts (e.g., number of interface interactions) or ratios (e.g., the conversion rate for the number of successful completions for a task divided by the number of task attempts) (Jansen & Clarke, 2017). These numbers are evaluated as key performance indicators (KPIs) (Maté et al., 2017), and in turn, organizations use these KPIs as measurements for progress toward achieving some objective or goal.

Companies employ user analytics for various purposes (Denizci Guillet, 2020; Salminen et al., 2020c; Thirumuruganathan et al., 2021, 2023), including system design, feature development, reputation management, CRM, marketing, advertising, cross-selling, upselling, and other related activities. For example, if the data is at the individual level, one can employ it to personalize interfaces or content. However, one often wants to aggregate the data to identify segments for tasks such as look-alike analysis or recommended features for systems. Following this, segmenting is the process of identifying user groups in the dataset with common behaviors, demographic attributes, or other unifying factors (Murray et al., 2017).

Although beneficial and actionable for many tasks, user analytics data is still a reasonable but remote proxy for the real users – that is, the actual users have been *de-personified* to numbers. The reliance on user analytics can lead to focusing on the numbers at the expense of a genuine user focus, user understanding, and user empathy. Since empathy has been shown to benefit design and user-centric decision making enabled by customer-oriented information systems (J. Iivari & Iivari, 2011; Iivari, 2009; Wechsler & Schweitzer, 2019) and is one of the qualities that computational social sciences researchers almost unanimously agree about (Jansen et al., 2020b; Pelau et al., 2021; Wright & McCarthy, 2008), personification techniques and personification artifacts are advantageous.

For this research, we leverage user analytics data and a persona analytics system to create one type of personification artifact, namely algorithmically-generated personas.

## 2.2 Algorithmically-Generated Personas

Personas are an example of personification. Personas depict fictional people created to represent real users (Cooper, 2004). In information systems research, personas have been deployed in association with multiple information system types, including recommender systems (Karumur et al., 2018; Yuan et al., 2013), human–robot interaction (Chien et al., 2022), information retrieval (Venkatsubramanian & Hill, 2010), public e-service systems (Holgersson et al., 2015), social network services (Choi et al., 2016), and so on. Personas are seen to facilitate the development of user-centric analytics information systems, but they can also be outcomes of the information system operations in themselves, i.e., serving as “surrogate users” to provide information for decision makers about the users of a given system, product, or service (Cooper, 2004).

Traditionally, algorithmically-generated personas have been created somewhat ad hoc from limited data sources (Nielsen, 2019), but the availability of large amounts of online user data has made creating personas more opportune (Jansen et al., 2021c). These

algorithmically-generated personas are created from user data using data science algorithms. This generation usually involves analyzing large amounts of real user data that is difficult or impossible to analyze manually for persona creation (Jansen et al., 2020b). As such, algorithmically-generated personas are acknowledged to be precise and accurate user representations via the personification of the data. As algorithmically-generated personas are often created via persona analytics systems (Mijač et al., 2018), these types of personas can be updated frequently so that they do not become stale or outdated. Therefore, algorithmically-generated personas provide humanized user representations that may be used to make better decisions about users and achieve a higher degree of user-centricity, which is expected to yield better business results (Jansen et al., 2020b).

The concept of algorithmically-generated personas has been presented in-depth in groundbreaking books (Cooper, 2004; Grudin & Pruitt, 2002) and later articles (An et al., 2018a, b; Nielsen, 2004; Pruitt & Grudin, 2003) on the subject and then extended by numerous researchers (Chang et al., 2008; Faily & Flechais, 2011; Nielsen et al., 2015). Algorithmically-generated personas have been investigated from several angles, including cultural (Meissner & Blake, 2011) and user segmentation representation (An et al., 2018a, b). Algorithmically-generated personas are more effective than depersonified analytics for certain user-focused tasks (Salminen et al., 2020b). Algorithmically-generated personas have also been shown to effectively present an accurate user representation (Jansen et al., 2020b) and rectify any incorrect preconceptions of users for stakeholders in organizations (Lauren Sorenson, 2011). As such, algorithmically-generated personas are an excellent instantiation of the personification process, and there are several published manuscripts on algorithmically-generated personas that study the various aspects of their creation and use in information systems and beyond (Jansen et al., 2021c; Mijač et al., 2018; Spiliotopoulos et al., 2020).

The research presented in this article leverages an ML model (Lee & Seung, 1999) and supporting algorithms (Blei et al., 2003; Darliansyah et al., 2019; Liu et al., 2020) to generate algorithmically-generated personas (Spiliotopoulos et al., 2020) from extensive user analytics datasets for personification. We investigate the effect of different hyperparameter values of the number of personas on the personification process in accurately representing the clusters (Zheng et al., 2022) of the underlying data.

## 2.3 Hyperparameters and their Effect on User Segmentation

Various ML approaches (Kelleher et al., 2020) use computing algorithms that ‘learn’ (i.e., improve performance) by



iterations using data. The approaches generally involve how much supervision the models are given or the specific ML models' use of statistical techniques (Celebi et al., 2016; Mohamed, 2017). Each ML approach has many alternative models that are particular instances of that approach. For example, one approach to ML is factorization, and one particular factorization model is non-negative matrix factorization (Lee & Seung, 1999), which is the ML model used in this research (due to its popularity in the field of data-driven persona generation; see, e.g., (An et al., 2018a, 2018b)). As user data availability, accessibility, and volume have increased, ML models are increasingly employed to improve user data analysis in information systems (Rust & Huang, 2014).

Most ML models have parameters that define the approach and the limits of the algorithmic analysis (Agrawal, 2020). Parameters are usually derived internally by the models from the data that the model uses. However, these ML models also have hyperparameters (Probst et al., 2009), which define the model's limits and usefulness for information systems (Thirumuruganathan et al., 2014), but these hyperparameters are not derived directly from the data. Hyperparameters are outside the data (Ibnu et al., 2019), usually determined by researchers from trial and error, domain knowledge, estimation, and/or rules of thumb. For example, clustering is often used to segment user analytics data (Ditton et al., 2021; Jansen et al., 2011; Wu & Chou, 2011). However, the 'right' number of clusters (Wu & Chou, 2011) is often a matter of opinion or related to some data-external organizational goal. Therefore, while most clustering algorithms require the number of clusters desired as a hyperparameter, selecting the appropriate hyperparameter requires tuning and experiments adapted to the information system.

The research presented here generates algorithmically-generated personas from user analytics data using an ML model as an example of personification. The hyperparameter of interest is the number of algorithmically-generated personas to create from the data. In this research, we evaluate the effect of the hyperparameter values on the representation by the underlying user analytics personification process, aiming towards making persona analytics systems more robust and helpful for decision makers facing an ever-increasing amount of user data.

### 3 Research Questions

Our core research question (RQ) is: *How does the hyperparameter of persona number affect the personification of user analytics data?* Regardless of the hyperparameter used to generate algorithmically-generated personas, one would want the resulting personas to accurately reflect the underlying user data. If we change the hyperparameter (i.e., a different number of personas), we need the resulting personas to be stable by accurately reflecting the underlying user data.

If different sets of personas give different views of the same set of users, these biased representations could adversely affect any decisions made by these personas. Addressing this RQ, we examine the effect of hyperparameter selection on four common demographic attributes of nearly all personas, saving the analysis of other attributes for other research. Specifically, we examine four hypotheses (H) concerning the personification bias:

- H1: Changing the number of personas alters the representation of user gender.
- H2: Changing the number of personas alters the representation of user age.
- H3: Changing the number of personas alters the representation of user nationality.
- H4: Changing the number of personas alters the representation of user gender-age-nationality (GAN).

Based on previous research on ML hyperparameters, there is reason to assume that varying the hyperparameter of persona number substantially affects the personification process, directly affecting the kinds of personas the algorithm produces. This premise is based on prior work in the ML area, the results of a pilot study (Jansen et al., 2021b), and combined with our experience in user and persona analytics systems. We operationalize our analysis with the quantification of personification bias, which we refer to as *the degree to which the results of the hyperparameter selection deviate from the baseline*. We also report conformity  $C$ , defined as *the degree of non-deviation from the baseline*. We report the deviation  $D$ , which equals  $D = 1 - C$ .

We select a range of hyperparameter values from 5 to 15 for the range of the data-driven persona creation algorithm that we employ in this research. The choice represents a three-time increase/decrease in the hyperparameter value, but which is still within the cognitive capabilities of the stakeholders who will have to employ personas, thus maintaining their value for decision making under bounded rationality (Simon, 1990). This is also a range that algorithmic persona studies commonly apply (see, e.g., An et al., 2018a, b; Blomquist & Arvola, 2002; Kim et al., 2019; Subrahmaniyan et al., 2018). Therefore, the tested hyperparameter range is realistic for actual employment and adheres to the commonly applied range seen in the literature.

We also choose a concatenation of gender, age, and nationality (GAN), as these demographic attributes are commonly used in the personification process when constructing persona profiles (Nielsen et al., 2015). In addition, gender, age, and nationality are standard attributes in many analytics information systems used in the industry (e.g., YouTube Analytics, Facebook Insights, Google Analytics). The combination of these attributes will also aid in accounting for skewed datasets. We note that the major analytics information systems use IP

location for nationality and biological sex as a proxy for gender, which is why the data follows the binary gender paradigm (male/female). We also investigate the concatenation of gender, age, and nationality (GAN), as this derived demographic attribute is commonly used in the personification process. The demographic variables have different inherent values ranging from reasonably small with gender (two options), limited with age (when condensed into seven age groupings, again, following the industry standard), and relatively large with nationalities.

Altogether, our choice of persona attributes provides a good spectrum for analysis. Furthermore, our research questions represent a span of demographic characteristics to evaluate the effect of hyperparameter value selection on the personification process for personifying user analytics data. As such, the research is of theoretical and practical value, especially when considering information systems that present user segments in one format or another.

## 4 Data Collection & Methods

### 4.1 Data Collection

We employ organizational user analytics data from three major online channels: Facebook (FB), Google Analytics (GA), and Instagram (IG). FB and IG are major social media services, and each provides account-level user analytics data via FB Insights and IG Insights. GA is the de facto industry user analytics service for websites. So, the three services are major online channels for companies to interact with their online user communities. Each service offers industry-standard analytics similar to many other major analytics information systems, such as Adobe Analytics, YouTube Analytics, IBM Analytics, and custom-made logging analytics. The individual user data is aggregated for personally identifying attributes or values for each of the three analytics information systems, so privacy concerns are minimal. The process employed in this research can be used with proprietary CRM data containing individual user information.

We collected user analytics statistics from these three analytics information systems for the focal organization, which is a major international news and content provider. The organization employs personas to monitor the reach and engagement of social media posts, news articles, videos, and so on. At the time of the study, the organization's FB account had more than 15.5 million followers, and its IG account had more than 2.3 million followers. The web traffic monitoring service

SimilarWeb reports more than 40.5 million monthly visits to the organization's website. The organizational accounts on each of the channels have thousands of posted content pieces and tens of millions of user interactions with this content. As such, the organization exemplifies modern enterprises with extensive and diverse online user populations, and the user analytics datasets employed in this research are extensive, heterogeneous, and correspond to real-world conditions.

We note that the user analytics data is available only to the channel account holders and not open to the general public. The data for this research is employed with the channel account holders' permission, who provided access to the data via an Application Programming Interface (API) to the persona analytics system described below.

### 4.2 Personification Process

We employ an industry-standard persona analytics system called Automatic Persona Generation (APG) for the personification process, available online at [URL MASKED FOR REVIEW]. APG takes large amounts of user analytics data and personifies the data by creating algorithmically-generated personas that aim to accurately and precisely represent the underlying user data. As the persona analytics system has been described in detail in other research (An et al., 2018a, b; Jansen et al., 2020a, b; Jung et al., 2017), we briefly present it here and refer the interested reader to published research. The personification is shown in Fig. 2.

APG leverages user analytics data from various possible sources, including CRM, FB, IG, FB Ads, Google Ads, GA, and YouTube, via accessing the APIs after inclusion in an organizational account of the particular channel. The user analytics data accessed by APG is aggregated information on demographics (e.g., gender, age group, country code) and behaviors (e.g., product id, count of user interactions). The behavioral data can be any product (e.g., webpage, ad, book) associated with one or more behaviors (e.g., visit, click, purchase). In the specific case of this research, the product is online content (for FB and IG) and webpages (for GA).

APG uses a non-negative matrix factorization (NMF) (Lee & Seung, 1999) algorithm as the initial step in personifying user analytics data. NMF is an ML factorization approach to identify sets of products related to user behaviors (e.g., sets of webpages interacted with by similar types of users). NMF builds a matrix of these products and



**Fig. 2** The APG personification approach is a six-step process to convert raw user analytics data into rich persona profiles

latent features via matrix decomposition. NMF also makes a matrix of demographic groups and their association with observed latent factors. NMF associates the demographic groups to the product sets using the latent factors, resulting in textual and numerical user profiles (i.e., skeletal personas). APG determines the gender, age, and nationality attributes from the data in the user analytics dataset via the service's analytics information system.

APG then enriches these skeletal user profiles to generate complete and rich algorithmically-generated personas (thus following the concept of rounded persona profiles from persona design theory (Nielsen, 2019)). APG employs an internal database of thousands of purchased stock photographs of real people, with each image manually tagged with an appropriate gender-age-nationality. The system also has an internal database of tens of thousands of names. Each name is also meta-tagged with an appropriate gender-age-nationality probabilistically to match the persona's gender-age-nationality, using a custom-built algorithm (Jung et al., 2021). Leveraging second- and third-party data from the FB Audience Manager and Twitter accounts, APG calculates the probability of other background information factors such as occupation, education, and relationship status. APG also determines the product topics of interest of the personas. APG generates these topics of interest (Jansen et al., 2019a) based on the products interacted with using a variety of classification algorithms, including zero-shot classification and supervised ML (Salminen et al., 2019).

### 4.3 Outcome of the Personification Process

Using various algorithmic approaches and periodic systemic data collection, APG determines the sentiment of social media comments in multiple languages, including English, Arabic, Turkish, Spanish, Finnish, and French, implemented using EmoLex (Mohammad & Turney, 2013). An example of a data-driven persona created by the APG persona analytics system is shown in Fig. 3, with the key features employed in this research annotated.

As a persona analytics system, APG has an advantage compared to manual approaches to persona creation. APG can generate a different number of personas in given persona sets or persona casts. Although theoretically, APG can create any number of personas, given the decision makers' cognitive limitation of managing a large number of personas (see Jansen et al., 2021a), the interface affords the generation of persona sets from 5 through 15 inclusive. The sets of personas are ranked by the size of the user segment they represent. As most organizations now have multiple online channels, APG offers a comparison feature to display sets of personas from multiple channels simultaneously, with an example of 5 personas from FB, GA, and IG shown in Fig. e 4.

### 4.4 Changing Hyperparameter Values

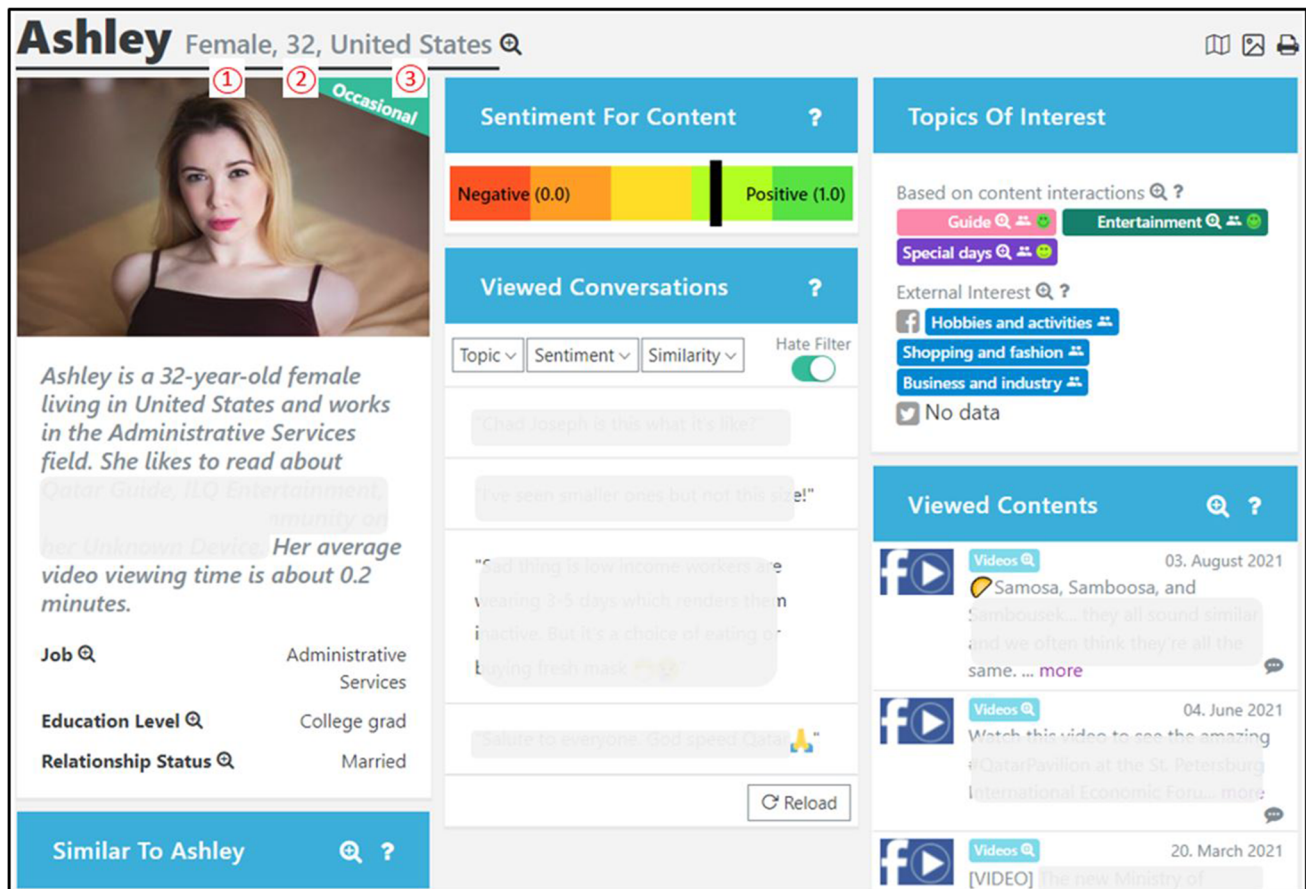
APG generates a new listing of algorithmically-generated personas for each hyperparameter selection. As the generation of personas from 5 to 15 is based on the *same* underlying user population data (i.e., the set of 5 personas is based on the same user interaction data as the set of 15 personas, for example), each of the persona listings is both independent (e.g., any of the personas in the set of 5 is autonomous from any persona in the set of 15 for example), regardless of similar demographic attributes and related behaviors. However, the same or similar personas can appear in multiple persona listings when the hyperparameter values change.

As the number of personas is increased from 5 to 15, the persona listing increases correspondingly, with the underlying user population being personified with greater granularity (see Table 1 in the results). As shown in Fig. 5, the hyperparameter (i.e., the number of personas generated) is increased from the five personas shown in Fig. 4 to ten in Fig. 5a and then fifteen in Fig. 5b.

Reviewing the persona sets in Figs. 3 and 4, we see that some portions of the persona listings intersect (i.e., the identical/similar personas are in multiple listings), and other personas are unique (i.e., the personas only appear in one persona listing). Also, as the focal persona analytics system segments the user analytics data based on the hyperparameter values, the ranking of some personas that may represent similar user segments across listings changes.

This ranking change results from the changing size and number of user segments represented by each set's personas. We also note that some personas and their underlying user segments are consistent across all personas listing and constant in ranking across all persona listings. For example, this applies to the two top-ranked personas, John (34 years old from the United States) and Raj (26 years old from India). Other personas appear in multiple listings but at different rankings, such as Priya (29 years old from India). In sum, a visual inspection of the persona listings indicates that the hyperparameters' change might skew the decision maker's perceptions of the user population when using a persona analytics system.

Whether 5, 15, or any number of personas in between, the total of the personas in the listing represents the *same* underlying user analytics dataset. Therefore, in a sense, each representation is 'correct,' as determined by the algorithm for that given hyperparameter value. However, as we need a baseline in order to compare across sets of personas, we generate all 11 persona sets (i.e., one set of personas that includes 5 personas, another with 6 personas, until 15 personas) and then calculate the proportional representation for gender, age, nationality, and GAN for this baseline. This gives us a 'gold standard' that we take as the



**Fig. 3** Example of an APG algorithmically-created persona profile with the three elements of the persona profile pertinent to this study annotated: (1) the persona’s gender, (2) the persona’s age, and (3)

the persona’s nationality. Organizational identifying information is masked with gray boxes

‘accurate’ representation based on majority vote thinking, i.e., the values obtained reflect the “collective view” of the algorithm across different hyperparameter values. This approach is similar to Jansen et al., (2019b, 2021a, b, c) where the number of personas varied in attempting to create manageable persona sets for decision makers relying on analytics information systems.

We then calculate the anticipated number of personas we expect in each hyperparameter setting with these percentages. For example, if the percentage for Male = 50% in the baseline, and the hyperparameter value is 6, then we would expect 3 of the personas to be male. We compare the *expected* number of personas to the *actual* number of personas that were generated in each set. We then determine if the actual number of personas is distorted – i.e., whether it exceeds (*Over*); is less than (*Under*), or is stable (*Same*) relative to the baseline. We can then determine whether the persona sets are stable or if they distort the representation by over or under-representing the user population for the given attribute. One can calculate the times a given

attribute is over or under, and in this way, quantify the “algorithmic bias” (Hajian et al., 2016) of hyperparameter setting in algorithmically-generated personas. Here, we are not arguing that this averaging approach is the *best* for determining the stability of personification because this is not the objective of our research, and any determination of the ‘best’ (if such a thing exists) we leave for other studies. However, the constructed baseline matches our research goal of determining if changing the hyperparameter values affects the personification process’s stability and/or distortion. As such, the baseline is sufficient for investigating our RQ.

For statistical analysis of the hypotheses, we employ the Wilcoxon matched pairs signed test (Ramsey et al., 1993), which is a statistical procedure that computes the difference between each set of matched pairs and then compares the sample median against a hypothetical median. We use the 33 hypermeter outcomes, comparing the actual results with the expected baseline values. We use the Wilcoxon test rather than the paired t-test because the data is not normally distributed.



**Table 1** Results of gender analysis of the 33 persona sets across the three channels, FB, GA, and IG. The set of 15 personas is the only hyperparameter with no bias. Cells with the highest values are shaded

Platform	Gender Attribute	Hyperparameter															Total	%	Deviation
		5	6	7	8	9	10	11	12	13	14	15							
FB	Same	0	0	2	2	0	2	2	0	2	2	2	14	63.6%	36.4%				
	Over	1	1	0	0	1	0	0	1	0	0	0	4	18.2%	18.2%				
	Under	1	1	0	0	1	0	0	1	0	0	0	4	18.2%	18.2%				
GA	Same	2	0	0	0	0	2	0	0	0	0	2	6	27.3%	72.7%				
	Over	0	1	1	1	1	0	1	1	1	1	0	8	36.4%	36.4%				
	Under	0	1	1	1	1	0	1	1	1	1	0	8	36.4%	36.4%				
IN	Same	2	0	0	0	0	0	2	0	0	0	2	6	27.3%	72.7%				
	Over	0	1	1	1	1	1	0	1	1	1	0	8	36.4%	36.4%				
	Under	0	1	1	1	1	1	0	1	1	1	0	8	36.4%	36.4%				
Same	Over	4	0	2	2	0	4	4	0	2	2	6	26	39.4%	30.3%				
	Under	1	3	2	2	3	1	1	3	2	2	0	20	30.3%	30.3%				
	Deviation	33.3%	100.0%	66.7%	66.7%	100.0%	33.3%	33.3%	100.0%	66.7%	66.7%	0.0%	60.6%						

## 5 Results

Here, we present the results of the analysis of varying the hyperparameter of the personas analytics system from sets of 5 to 15 personas based on the accuracy outcomes of gender, nationality, and topics of interest. We begin with the results from a primary analysis of the personas generated from the user population dataset and continue with a confirmatory analysis of datasets independent from those used in the primary analysis.

### 5.1 Primary Analysis
















We analyzed the 330 personas generated from the 11 generations of the persona sets, from 5 to 15 personas inclusive, for the organization for each of the three channels. Among these were 255 male personas (68.2%) and 75 female personas (31.8%). To our knowledge, all of the analytics information systems that we are aware of use biological sex as a proxy for gender identity. However, we expect this to change in the future as more nuanced metrics are incorporated into these user analytics information systems. So, we reserve any analysis for other gender identities for future research.

The average age for the 330 personas was 31.9 years (SD = 11.23, max = 74, min = 18, med = 31), reflecting the generally younger user population of the online channels. For the analysis, we clustered the 330 personas into age groupings so as to reflect the age clusters in the underlying analytics information system, generally mirroring the US Census categories and resulting in seven age groupings for the 330 personas. The most common age groupings are the 25–34 age category (66.7% of the personas), the 18–24 age category (15.8%), and the 35–44 age category (8.2%). The 110 personas are from 14 countries, representing 7.1% of the world's 198 countries. The most commonly occurring countries were the United States (28.5%), India (21.5%), and the United Kingdom (10.3%).

Concerning GAN, there are theoretically 2,772 possibilities (i.e., 2 genders  $\times$  7 age-categories  $\times$  198 countries). In our collective datasets, there were 38 unique GAN combinations (1.4% of the theoretical limit), with the following being the most common: [Male, 25–34, India] (15.2%), [Male, 25–34, United States] (8.5%), and [Female, 25–34, United States] (7.3%). We now move to the analysis addressing our RQ concerning gender, age, nationality, and GAN.































#### 5.1.1 Gender

Concerning H1: *Changing the number of personas alters the representation of user gender*; ten of the eleven persona sets present a biased representation of the user population,

Facebook	Google Analytics	Instagram
 <b>John</b> 34, United States	 <b>Michael</b> 34, United States	 <b>Rahul</b> 31, India
 <b>Raj</b> 26, India	 <b>James</b> 60, United States	 <b>Usman</b> 25, Pakistan
 <b>Amanda</b> 30, United States	 <b>John</b> 69, United States	 <b>Ashley</b> 33, United States
 <b>Rahul</b> 22, India	 <b>Abhishek</b> 32, India	 <b>John</b> 28, United States
 <b>Aldrin</b> 29, Philippines	 <b>Issam</b> 32, Qatar	 <b>Putri</b> 20, Indonesia

**Fig. 4** Example of the APG persona analytics system interface, with the cast of personas (in this case, a set of 5 personas) from three online channels. The personas in the listing are by default ranked by the current reach of the personas, which is based on the persona's rep-

resentativeness of the baseline user data. Clicking on a result in the persona listing displays the corresponding persona profile (example displayed in Fig. 2). Note: The purple bars at the top have organizational and date-stamped information that has been masked

Facebook	Google Analytics	Instagram
 <b>John</b> 34, United States	 <b>Michael</b> 34, United States	 <b>Chris</b> 29, United States
 <b>Raj</b> 26, India	 <b>John</b> 69, United States	 <b>Abhishek</b> 32, India
 <b>Chris</b> 42, United States	 <b>Abhishek</b> 32, India	 <b>Jason</b> 41, United States
 <b>Amanda</b> 30, United States	 <b>Justin</b> 27, Canada	 <b>Rahul</b> 22, India
 <b>James</b> 25, United Kingdom	 <b>Jessica</b> 28, United States	 <b>James</b> 48, United States
 <b>Rahul</b> 22, India	 <b>Paul</b> 37, United Kingdom	 <b>Ashley</b> 33, United States
 <b>Usman</b> 27, Pakistan	 <b>Herb</b> 74, Canada	 <b>Asad</b> 19, Pakistan
 <b>Aldrin</b> 29, Philippines	 <b>Mark</b> 45, United Kingdom	 <b>Aldrin</b> 33, Philippines
 <b>Tanvir</b> 33, Bangladesh	 <b>Issam</b> 46, Nigeria	 <b>Sakib</b> 24, Bangladesh
 <b>Priya</b> 30, India	 <b>Segun</b> 46, Nigeria	 <b>Rogers</b> 29, Uganda

(a)
(b)

**Fig. 5** Value of the hyperparameter — i.e., number of algorithmically-generated personas generated from user analytics data by APG — for 10 (a) and 15 (b) personas, with the corresponding persona

listing for each. The listings for 10 and 15 are shown as examples. The APG system can generate casts of personas from 5 to 15, inclusive

ranging from 33.3% to 100.0%, with an overall deviation of 52.0%. Table 1 shows the platforms in column 1, the outcomes of the hyper parameter changes (i.e., ‘same’, ‘over’, ‘under’) in Column 2, and the counts in each row with the specific hyperparameter in the corresponding columns.

However, although deviation (and conformity) varied among the channels, there were notable trends. The deviation was notably lower on FB (36.4%) compared to GA (72.7%) and IG (72.7%). A Wilcoxon matched pairs signed test was conducted to determine whether there was a difference in the actual distribution for the gender of the generated personas compared to the expected values. The analysis results show a significant deviation for gender,  $z = 5.0119$ ,  $p < 0.01$ , indicating a bias in personification. Therefore, H1 is fully supported; *the number of personas alters the representation of user gender*.

### 5.1.2 Age

Concerning H2: *Changing the number of personas alters the representation of user age*, as shown in Table 2, all eleven of the persona sets have some level of bias, ranging from 27.8% to 71.2%, with 51.0% being the average. The hyperparameter of 10 personas provided the highest stability. Deviation again varied among the channels, although lower relative to gender distortion. Distortion was within ten percentage points for all channels, FB (51.5%), GA (55.6%), and IG (42.4%). Addressing H2, a Wilcoxon matched pairs signed test was conducted to determine whether there was a difference in actual distribution for the age of the generated personas compared to the expected distribution. The analysis results show a significant deviation for age in the actual personas,  $z = -4.703$ ,  $p < 0.01$ , indicating a bias in

personification. Therefore, H2 is fully supported; *the number of personas alters the representation of user age*.

### 5.1.3 Nationality

For H3: *Changing the number of personas alters the representation of user nationality*; as shown in Table 3, none of the eleven persona sets were bias-free, although persona set 8 was close, with only a 4.8% deviation. The other persona sets had deviations ranging from 14.3% to 33.3%. (79.5%). Overall, all of the persona sets had reasonably consistent representations, with an average deviation of 22.9%. Distortion (and conformity) varied somewhat among the channels, with FB (13.0%) and IG (16.9%) below that of GA (39.0%). Addressing H3, a Wilcoxon matched pairs signed test was conducted to determine whether there was a difference in actual distribution for the nationality of the generated personas compared to the expected distribution. The analysis results show a significant deviation for nationality in the actual personas,  $z = -4.1069$ ,  $p < 0.01$ , indicating a bias in personification. Therefore, H3 is fully supported; *the number of personas alters the representation of user nationality*.

**5.1.3.1 GAN** Examining topics of interest for H4: *Changing the number of personas alters the representation of user GAN*; as shown in Table 4, all eleven of the persona sets had some level of biased representation, with an average deviation of 40.0%. Distortion (and conformity) varied somewhat among the channels: FB (11.2%) and IG (18.2%), with GA notably higher (74.8%). Addressing H4, a Wilcoxon matched pairs signed test was conducted to determine whether there was a difference in actual distribution for GAN of the generated

**Table 2** Results of age analysis of the 99 persona sets for three organizations across the three channels: FB, GA, and IG. Set 10 has the best representation of the user population. Cells with the highest values are shaded

Platform	Age Attribute	Hyperparameter											Total	%	Deviation
		5	6	7	8	9	10	11	12	13	14	15			
FB	Same	4	2	2	2	4	4	4	4	2	2	2	32	48.5%	51.5%
	Over	1	2	2	2	1	1	1	1	2	2	2	17	25.8%	
	Under	1	2	2	2	1	1	1	1	2	2	2	17	25.8%	
GA	Same	4	4	4	4	2	6	4	6	4	4	2	44	44.4%	55.6%
	Over	2	2	2	2	3	1	2	1	2	2	3	22	22.2%	
	Under	3	3	3	3	4	2	3	2	3	3	4	33	33.3%	
IN	Same	1	1	1	3	3	3	3	1	1	1	1	19	57.6%	42.4%
	Over	1	1	1	0	0	0	0	1	1	1	1	7	21.2%	
	Under	1	1	1	0	0	0	0	1	1	1	1	7	21.2%	
Same		9	7	7	9	9	13	11	11	7	7	5	95	48.0%	
Over		4	5	5	4	4	2	3	3	5	5	6	46	23.2%	
Under		5	6	6	5	5	3	4	4	6	6	7	57	28.8%	
Deviation		50.0%	61.1%	61.1%	50.0%	50.0%	27.8%	38.9%	38.9%	61.1%	61.1%	72.2%	52.0%		

**Table 3** Results of nationality analysis of the 99 persona sets for three organizations across the three channels: FB, GA, and IG. Set 8 has the best representation of the user population. Cells with the highest values are shaded

Platform	Nation-ality Attribute	Hyperparameter												Total	%	Deviation
		5	6	7	8	9	10	11	12	13	14	15				
FB	Same	7	4	5	6	7	6	7	6	7	7	5	67	87.0%	13.0%	
	Under	0	2	1	1	0	0	0	0	0	0	1	5	6.5%		
	Over	0	1	1	0	0	1	0	1	0	0	1	5	6.5%		
GA	Same	3	4	4	7	5	5	5	3	3	5	3	47	61.0%	39.0%	
	Under	2	1	1	0	1	1	1	2	2	1	2	14	18.2%		
	Over	2	2	2	0	1	1	1	2	2	1	2	16	20.8%		
IN	Same	6	7	7	7	5	7	5	5	5	3	7	64	83.1%	16.9%	
	Under	0	0	0	0	1	0	1	1	1	2	0	6	7.8%		
	Over	1	0	0	0	1	0	1	1	1	1	2	0	7		9.1%
Same		16	15	16	20	17	18	17	14	15	15	15	178	77.1%		
Over		2	3	2	1	2	1	2	3	3	3	3	25	10.8%		
Under		3	3	3	0	2	2	2	4	3	3	3	28	12.1%		
Deviation		23.8%	28.6%	23.8%	4.8%	19.0%	14.3%	19.0%	33.3%	28.6%	28.6%	28.6%	22.9%			

personas compared to the expected distribution. The analysis results show a significant deviation for GAN in the actual personas,  $z = -4.703$ ,  $p < 0.01$ , indicating a bias in personification. Therefore, H4 is fully supported; *the number of personas alters the representation of user gender-age-nationality*.

The analysis above confirms that tuning the value of hyperparameters for personification substantially affects the user data representation.

## 5.2 Confirmatory Analysis

We conducted the above analysis for two other organizations to evaluate the robustness of our findings using independent

datasets. We conducted the analysis using the same procedure outlined above for the YouTube channels of two other organizations (ORG2 and ORG3), briefly described below. The names are masked for the anonymity of the organizations. These organizations operate in different industries, and their range of content production and online presence varies from small (less than a million followers) to large (millions of followers). Therefore, testing the hyperparameters on these different datasets will enable us to evaluate if the results are consistent across different organizational contexts.

- **ORG2:** a large non-profit organization promoting education, research, and community development to

**Table 4** Results of GAN analysis of the 99 persona sets for three organizations across the three channels: FB, GA, and IG. Set 6 is best representation of the user population. Cells with the highest values are shaded

Platform	GAN Attribute	Hyperparameter												Total	%	Deviation
		5	6	7	8	9	10	11	12	13	14	15				
FB	Same	11	13	11	13	11	11	11	11	13	13	9	127	88.8%	11.2%	
	Under	1	0	1	0	1	1	1	1	0	0	2	8	5.6%		
	Over	1	0	1	0	1	1	1	1	0	0	2	8	5.6%		
GA	Same	8	7	5	6	5	6	5	4	6	4	5	61	25.2%	74.8%	
	Under	0	1	1	1	2	1	2	2	1	2	2	15	6.2%		
	Over	14	14	16	15	15	15	15	16	15	16	15	166	68.6%		
IN	Same	16	16	16	14	14	16	16	14	14	16	10	162	81.8%	18.2%	
	Under	1	1	1	2	2	1	1	2	2	1	4	18	9.1%		
	Over	1	1	1	2	2	1	1	2	2	1	4	18	9.1%		
Same		35	36	32	33	30	33	32	29	33	33	24	350	60.0%		
Over		2	2	3	3	5	3	4	5	3	3	8	41	7.0%		
Under		16	15	18	17	18	17	17	19	17	17	21	192	32.9%		
Deviation		34.0%	32.1%	39.6%	37.7%	43.4%	37.7%	39.6%	45.3%	37.7%	37.7%	54.7%	40.0%			



a worldwide audience. The organization uses personas to better understand its online social media and other users and for strategic planning, involving crafting agendas to better serve the organization's various stakeholder groups. At the time of the study, the organization's FB account had more than 332,692 followers, and its IG account had more than 224,000 followers. SimilarWeb (<https://www.similarweb.com>—an industry-standard website competitive intelligence service) reports the organization's monthly website visits to be more than 106,000.

- **ORG3:** a small medium-sized enterprise (SME) providing tourism and lifestyle content concerning events, entertainment, dining, sports, and culture. In pursuit of this objective, the organization uses personas to better understand its online users in order to increase engagement and expand its audience to, hopefully, increase revenue. At the time of the study, the organization's FB account had more than 332,586 followers, and its IG account had more than 98,000 followers. SimilarWeb reports that the organization's website has more than 334,500 monthly visits.

Again, as shown in Tables 5, 6, 7, and 8, the analysis confirms that tuning the value of hyperparameters for personification substantially affects the user data representation.

## 6 Discussion

### 6.1 General Discussion

This research examines the effect of changing the hyperparameter values of a persona analytics system through eleven iterations of values from 5 to 15 personas, inclusive, across three major analytics information systems: FB, GA, and IG. The results show that changing the number of personas during the personification biases the representation of all four of the user analytics attributes that we measured, which were gender, age, nationality, and a combination of the three. For channels in general, representativeness is best on IG and FB for all organizations and all persona attributes, and bias is highest on GA for all organizations and all persona attributes. Although this requires further investigation, it seems reasonable that GA, representing heterogeneous web traffic, would have a more heterogeneous user population than social media channels. All four hypotheses were supported by statistical analysis. The highlights are as follows:

- Concerning *gender*, there was a predominance of males (68.2%) in the user population, but this did not appear to impact the hyperparameter selection effects in the per-

sonification process. Male personas and female personas were distorted (either over or under) at equal rates. For the 330 personas in the three personas sets from three channels, the sets were correct representations 39.4% of the time and biased (either under- or over-representation) 60.6% of the time (see Table 1). There were channel-specific differences, with FB having a substantially lower distortion (36.4%). The most representative persona set was persona set 15, which perfectly conformed to the baseline.

- Concerning *age grouping*, the 25–34 age category represented more than half (66.7%) of the baseline age groupings, with a persona set of 10 having the lowest deviation. The hyperparameter selection effect on age was less pronounced relative to gender, with 48.0% of the persona sets confirming with the baseline. The deviations were also similar across the three channels, ranging from a low of 42.4% to a high of 55.6% (see Table 2).
- Concerning *nationality*, the changing of the hyperparameters biased the personification representation of the user population. However, the distortion less drastic than with gender or age, with an average conformity of 71.1%. The highest deviation was 33.3% for the 15 personas set. Overall, of the 330 personas, the representative deviation was 22.9%. Distortion on GA was the highest (39.0%), lowest on FB (13.0%), and also low on IG (16.9%) (see Table 3).
- Moving to *GAN*, changing the personification hyperparameters resulted in biasing the representation of the user population. The hyperparameter setting of 6 had the lowest deviation (33.1%), but there were other persona sets with nearly as good results (i.e., 5 (34.0%). The overall conformity of the GAN attributes was relatively good (60.0%), being higher than gender (39.4%) and age (48.0%) and lower than nationality (77.1%). This conformity score seems reasonable, as GAN combines the other three attributes. However, the result is also surprising as we expected that deviation would be higher due to the 'curse of dimensionality' prevalent in ML datasets (Huang et al., 2019). We surmise that the result is due to the NMF approach used for the personification process and the structure of the underlying demographic attributes in major analytics information systems, but this premise needs to be investigated in future research.

Figure 6 compares the conformity and deviation (both over and under) for each hyperparameter for overall trends for all four attributes for the three organizations. The total distortion from hyperparameter selection is 39.0%, meaning that 61.0% of the representation is stable across hyperparameters. As shown in Fig. 6, perhaps

**Table 5** Results of gender analysis of the 99 persona sets for two organizations across the three channels: FB, GA, and IG. Persona sets 9, 10, and 11 are the best representations of the user population. Cells with the highest values are shaded

Platform	Org	Gender Attribute	Hyperparameter												Total	%	Deviation
			5	6	7	8	9	10	11	12	13	14	15				
FB	ORG02	Same	0	2	2	2	2	0	2	2	0	2	0	14	63.6%	36.4%	
		Over	1	0	0	0	0	1	0	0	0	1	0	4	18.2%		
	ORG03	Under	1	0	0	0	0	1	0	0	0	1	0	4	18.2%		
		Same	2	2	2	2	2	2	2	2	0	2	2	18	81.8%	18.2%	
		Over	0	0	0	0	0	0	0	0	1	0	0	2	9.1%		
		Under	0	0	0	0	0	0	0	0	1	0	0	2	9.1%		
GA	ORG02	Same	0	0	0	0	0	2	0	0	0	0	0	2	9.1%	90.9%	
		Over	1	1	1	1	1	0	1	1	1	1	1	10	45.5%		
		Under	1	1	1	1	1	0	1	1	1	1	1	10	45.5%		
		Same	2	0	2	0	2	2	0	2	2	2	0	14	63.6%	36.4%	
		Over	0	1	0	1	0	0	1	0	0	0	1	4	18.2%		
		Under	0	1	0	1	0	0	1	0	0	0	1	4	18.2%		
IN	ORG02	Same	0	0	0	0	2	2	2	0	2	0	0	6	27.3%	72.7%	
		Over	1	1	1	1	0	0	0	0	1	1	1	8	36.4%		
		Under	1	1	1	1	0	0	0	0	1	1	1	8	36.4%		
		Same	2	2	0	0	2	0	2	2	2	0	0	10	45.5%	54.5%	
		Over	0	0	1	1	0	1	0	0	0	1	1	6	27.3%		
		Under	0	0	1	1	0	1	0	0	0	1	1	6	27.3%		
Same		Over	0	2	2	2	4	4	4	4	4	2	0	22	33.3%		
		Under	3	2	2	2	1	1	1	2	3	2	3	22	33.3%		
	Deviation	Over	3	2	2	2	1	1	1	2	3	2	3	22	33.3%		
		Under	3	2	2	2	1	1	1	2	3	2	3	22	33.3%		
			100.0%	66.7%	66.7%	66.7%	33.3%	33.3%	33.3%	66.7%	100.0%	66.7%	100.0%	66.7%			

**Table 6** Results of age analysis of the 99 persona sets for two organizations across the three channels: FB, GA, and IG. Set 5 is the best representation of the user population. Cells with the highest values are shaded

Platform	Org	Age Attribute	Hyperparameter												Total	%	Deviation
			5	6	7	8	9	10	11	12	13	14	15				
FB	ORG02	Same	6	4	4	3	3	4	4	4	4	4	4	44	66.7%	33.3%	
		Over	0	1	1	1	1	1	1	1	1	1	1	10	15.2%		
		Under	0	1	1	2	2	1	1	1	1	1	1	12	18.2%		
	ORG03	Same	3	3	3	3	5	5	3	3	5	3	3	39	59.1%	40.9%	
		Over	2	2	2	2	1	1	2	2	1	2	2	19	28.8%		
		Under	1	1	1	1	0	0	1	1	0	1	1	8	12.1%		
GA	ORG02	Same	7	7	5	7	7	7	5	7	5	7	7	73	73.7%	26.3%	
		Over	1	1	2	1	1	1	2	1	1	1	1	13	13.1%		
		Under	1	1	2	1	1	1	2	1	1	1	1	13	13.1%		
	ORG03	Same	5	5	5	5	7	5	3	3	3	5	5	51	51.5%	48.5%	
		Over	2	2	2	2	1	2	3	3	3	2	2	24	24.2%		
		Under	2	2	2	2	1	2	3	3	3	2	2	24	24.2%		
IN	ORG02	Same	2	2	1	1	1	2	1	1	1	1	2	16	72.7%	27.3%	
		Over	0	0	1	1	1	0	1	1	0	1	0	6	27.3%		
		Under	0	0	0	0	0	0	0	0	0	0	0	0	0.0%		
	ORG03	Same	1	1	1	3	1	3	1	3	1	0	0	15	45.5%	54.5%	
		Over	1	1	1	0	1	0	1	0	1	2	1	9	27.3%		
		Under	1	1	1	0	1	0	1	0	1	1	2	9	27.3%		
Same		15	13	10	11	11	13	10	12	13	12	13	133	71.1%			
	Over	1	2	4	3	3	2	4	3	2	3	2	29	15.5%			
	Under	1	2	3	3	3	2	3	2	2	2	2	25	13.4%			
	Deviation	11.8%	23.5%	41.2%	35.3%	35.3%	23.5%	41.2%	29.4%	23.5%	29.4%	23.5%	28.9%				

**Table 7** Results of nationality analysis of the 99 persona sets for two organizations across the three channels: FB, GA, and IG. Set 5 is the most stable representation of the user population. Cells with the highest values are shaded

Platform	Org	Nationality Attribute	Hyperparameter												Total	%	Deviation
			5	6	7	8	9	10	11	12	13	14	15				
FB	ORG02	Same	5	5	7	5	5	5	3	5	5	3	3	51	66.2%	33.8%	
		Over	1	1	0	1	1	1	2	1	1	1	2	13	16.9%		
		Under	1	1	0	1	1	1	2	1	1	1	2	13	16.9%		
	ORG03	Same	6	6	4	4	4	4	2	4	4	2	6	50	64.9%	35.1%	
		Under	1	1	2	2	2	2	3	2	2	1	1	19	24.7%		
		Over	0	0	1	1	1	1	2	1	1	0	0	8	10.4%		
GA	ORG02	Same	7	7	5	7	7	7	5	7	7	5	7	73	66.4%	33.6%	
		Under	1	1	2	1	1	1	2	1	1	1	1	13	11.8%		
		Over	1	1	2	1	1	1	2	1	1	1	1	13	11.8%		
	ORG03	Same	5	7	5	4	5	5	7	5	4	5	4	56	50.9%	49.1%	
		Under	2	1	2	2	2	2	1	2	3	2	3	22	20.0%		
		Over	3	2	3	4	3	3	2	3	3	3	3	32	29.1%		
IN	ORG02	Same	11	9	7	7	11	9	9	9	7	9	9	95	77.9%	22.1%	
		Under	0	1	2	2	0	1	1	2	1	1	2	13	10.7%		
		Over	0	1	2	2	0	1	1	2	1	1	2	13	10.7%		
	ORG03	Same	8	6	6	6	8	8	6	4	6	6	4	68	77.3%	22.7%	
		Under	0	1	1	1	0	0	1	2	1	1	2	10	11.4%		
		Over	0	1	1	1	0	0	1	2	1	1	2	10	11.4%		
Same			23	21	19	19	23	21	17	19	21	19	219	73.7%			
Over			2	3	4	4	2	3	5	4	3	4	5	39	13.1%		
Under			2	3	4	4	2	3	5	4	3	4	5	39	13.1%		
Deviation			14.8%	22.2%	29.6%	29.6%	14.8%	22.2%	37.0%	29.6%	22.2%	29.6%	37.0%	26.3%			



**Table 8** Results of GAN analysis of the 99 persona sets for two organizations across the three channels: FB, GA, and IG. Set 5 is the most stable representation of the user population. Cells with the highest values are shaded

Platform	Org	GAN Attribute	Hyperparameter												Total	%	Distortion
			5	6	7	8	9	10	11	12	13	14	15				
FB	ORG02	Same	11	9	7	7	7	7	5	9	9	13	9	11	97	46.4%	53.6%
		Over	4	5	6	6	6	7	5	5	5	3	5	4	56	26.8%	
		Under	4	5	6	6	6	7	5	5	5	3	5	4	56	26.8%	
	ORG03	Same	9	7	7	7	5	9	9	9	7	9	8	7	84	44.9%	55.1%
		Under	4	5	5	5	6	4	4	4	5	4	5	5	52	27.8%	
		Over	4	5	5	5	6	4	4	4	5	4	4	5	51	27.3%	
GA	ORG02	Same	16	14	14	10	12	10	12	14	14	12	14	14	142	53.8%	46.2%
		Under	4	5	5	7	6	7	6	6	5	6	5	5	61	23.1%	
		Over	4	5	5	7	6	7	6	6	5	6	5	5	61	23.1%	
	ORG03	Same	20	20	18	18	20	20	18	18	16	16	16	16	200	90.9%	9.1%
		Under	0	0	1	1	0	0	1	1	2	2	2	2	10	4.5%	
		Over	0	0	1	1	0	0	1	1	2	2	2	2	10	4.5%	
IN	ORG02	Same	11	11	11	9	9	11	7	7	7	11	13	13	113	60.4%	39.6%
		Under	3	3	3	4	4	3	5	5	3	3	2	2	37	19.8%	
		Over	3	3	3	4	4	3	5	5	3	3	2	2	37	19.8%	
	ORG03	Same	11	11	9	13	11	11	11	11	11	11	11	11	121	57.9%	42.1%
		Under	4	4	5	3	4	4	4	4	4	4	4	4	44	21.1%	
		Over	4	4	5	3	4	4	4	4	4	4	4	4	44	21.1%	
Same Over Under Distortion			30	26	24	21	23	18	26	28	28	25	27	276	41.3%		
			11	13	14	17	16	17	16	16	15	12	12	11	154	23.1%	
			19	21	20	26	23	25	22	22	21	20	21	20	238	35.6%	
			50.0%	56.7%	58.6%	67.2%	62.9%	70.0%	59.4%	56.3%	53.3%	56.9%	53.4%	58.7%			

counter-intuitively, a high number of personas does not better represent the user population. Instead, it may lead to bias from over or under representation (Salminen et al., 2020). This finding is important because, intuitively, one would presume that increasing the number of personas improves the data representation in user segmentation. As with many other ML and algorithmic models, the hyperparameters must be tuned to the specific dataset.

## 6.2 Theoretical implications

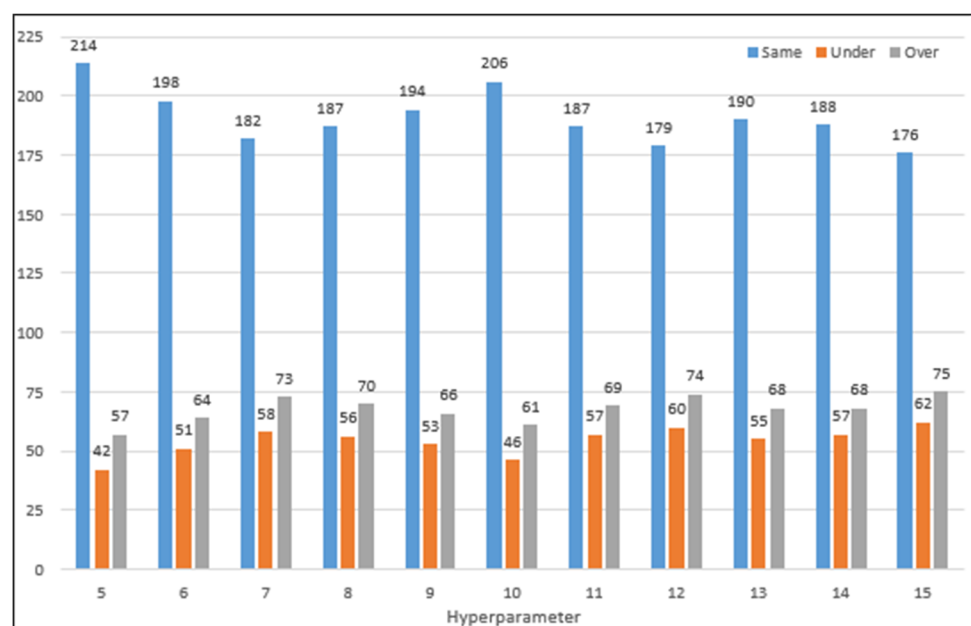
Concerning our core RQ (*How does the hyperparameter of persona number affect the personification of user analytics data?*), the presence of an effect is clear and manifests in the study results. Of the four persona attributes studied, in all four cases, the hyperparameter selection during the personification process resulted in a distorted representation of the user population to varying degrees. The results underline the impact of user-selected hyperparameters on the personification process. Specifically, the findings show that the number of personas generated can distort the attribute distribution of the generated personas, an example of the interplay between users and algorithmically-generated personas. It is also apparent from the analysis presented above that for a precise representation of users during personification, *the hyperparameters should be tuned to the specific dataset*. In other words, it might not be possible to recommend a generally universal “magic number” of personas for users of analytics information systems.

This finding indicates explicitly that ML personification approaches employing big data from user analytics

to create human-like representations and user segments, such as personas, need to expend the effort to tune the ML model to the appropriate hyperparameters. This tuning will determine the number of personas that will provide the most accurate representation of the underlying user data for the organization task at hand. Concerning the size of the user population, the larger the user population, the higher the bias. Again, this seems reasonable, as one would expect, and generally, the heterogeneity of the user population would be correlated with the size of the user population. However, this is mitigated somewhat by the channel effect, where sizeable user populations can be more homogenous. Finally, each persona attribute has a possible range of values (e.g., a few for gender, many for nationality). Generally, the more values in an attribute, the more bias in the personalization. However, again, this appears somewhat mitigated by the channel and homogenous populations. These factors open the door to aspects of theoretical tuning of the approach hyperparameters (i.e., number of personas) via automated means by entering a range of user population factors such as size and homogeneity. Determining this ‘theoretical tuning’ needs to be addressed in future research.

Our findings are also somewhat surprising because our premise was that a higher number of personas would categorically be more accurate, providing a better representation of the user population, as is presumed in the literature (Tang et al., 2006). In our analysis, this premise did not hold. Generally, increasing the size of the personas set did not increase compatibility with the baseline values, with often the 5-persona set having the highest conformity. The compatibility with the baseline

**Fig. 6** The effect of hyperparameter selection on the total of the 3,443 algorithmically-created persona attributes, as measured by *Same* (i.e., identical with baseline), *Over* (i.e., over the baseline), and *Under* (i.e., under the baseline). Interestingly, the 15-persona set, which one would expect to be the most granular, has the lower conformity



values of representation diminishes with an increase in the hyperparameter, with the set of 15 personas often being among the least accurate representations. While it has been noted in prior work (Chapman & Milham, 2006; Chapman et al., 2008) that increasing the number of persona attributes decreases the representativeness (Nasraoui et al., 2007), given that the personas sets are created from the same analytics set and contain the same attributes, we still find this outcome surprising.

A potential explanation for this could be the small number of categories for gender (2 categories) and age (7 categories), along with the average for the baseline. This “flaw of averages” is well known and documented, e.g., a classic study in 1950 conducted by the US Air Force, found that among 4,000 measured pilots, no pilots matched all of the average attributes of height, weight, etc. (Hertzberg et al., 1954). As a corollary from this flaw of averages, no set of user representations may exactly match the real composition. Nonetheless, the flaw of averages pertains to all customer and user segmentation efforts in analytics information systems, since these data-driven methods aim to aggregate information about the user population to a denser form of presentation.

### 6.3 Practical implications

Concerning the practical implications resulting from this research, we discuss the following three points that strike us as important.

**Insights from Trends in Personification Biases** There are trends in the degree of personification bias with an interplay among channels, size of the user population, and the range of attribute values, with some outliers in each. From these trends, there are at least two insights for the practical personification of user analytics data, which are: (1) hyperparameters, which as in most ML models, must be tuned to the specific dataset; and (2) as a rule of thumb to mitigate bias, the number of personas is positively correlated with both the number of attribute values and the size of the user population – the more attribute categories, the more personas; the larger the user population, the more personas. Modern social media platforms and websites can have user populations in the millions, indicating sets of personas beyond the historical ‘handful’ advocated in prior literature (Salminen et al., 2022), and perhaps hundreds of personas. Persona sets of these numbers necessitate the need for interactive persona systems that offer searching, filtering, sharing, and collaboration on options for decision makers using a persona analytics system. The study of this specific hyperparameter, in the context of statistical analysis for organizations, can help determine better personas for more effective marketing strategies

and understanding of the target demographics of specific brands/products.

**Avoiding the Mystique of Numbers** When employing advanced personification techniques, such as algorithmically-generated personas, in analytics information systems, decision makers need to be aware that a set of personas may give a biased view of the user population (Laporte et al., 2012; Siegel, 2010), at least in terms of gender, age, nationality, and GAN, as shown in this research. Just because the obtained ‘answer’ involves a lot of data and an algorithm, it does not necessarily mean the ‘answer’ is accurate or the only possible way of perceiving the data. This ideal requires that developers of algorithmically-generated personas and other personification techniques implemented in analytics information systems adequately execute the tuning of the ML models, and perhaps report results obtained using different hyperparameters. Decision makers using a persona analytics system can benefit from adjusting the model hyperparameters and observing the concrete results of their changes. In other words, decision-makers using persona analytics systems must understand hyperparameter tuning before taking action on the personas.

**Personification of User Data Needs an ‘X’** It may be inappropriate to rely on the data by itself to determine the ML hyperparameter selection if the overall goal (e.g., design criteria, business decision, engagement objective) is external to the user analytics data or the ML model. In this research, we were interested in a representation of the users. However, there may be other business goals where the representation of the user population is not appropriate. For example, an organization may want to emphasize the diversity within a user population (Sheth et al., 2000) to highlight small segments for targeting or emerging segments. Organizations may wish to represent the most impactful user segments (Reinartz & Kumar, 2000) or the least costly user segments (Helgesen, 2006). A more extensive set of personas would seem more appropriate for cases. These numerical outliers are consumed or hidden within the ‘average’ of the overall user population characteristics for many algorithmic approaches. Thus, appropriate hyperparameter selection in these external organizational cases is an important area for future research.

### 6.4 Strengths, Limitations, and Future Work

As with most research, there are both strengths and limitations. For strengths, we employed multiple large user analytics datasets numbering in the tens of millions of

user actions from three different organizations of various user population sizes, across three major online services employed by thousands of organizations. So, the datasets are representative of those entities with extensive and diverse user populations, implying that the overall approach is generalizable. The personification process of this user data was accomplished using a state-of-the-art persona analytics system employing a robust factorization approach for personification of the analytics data, and the process was validated across 9 datasets.

Concerning limitations, we leveraged only one personification approach, that of algorithmically-generated personas. Other forms of personification and user modeling, such as scenarios or user profiles, should be examined to see how ML models in these areas are affected by hyperparameter settings. However, given that the models are standard across personification approaches, one would expect the trends shown in these research findings to hold with these other personification approaches. However, this would need to be confirmed in future work.

Regarding other limitations, we employed nine data sets, one ML model, one metric (i.e., stability), four attributes (i.e., gender, age, nationality, GAN), and one baseline (i.e., an average of the personification sets). Future work should explore the setting of personification hyperparameters on other user data sets from other companies, other user data types (e.g., retail data or user relations management data), and other analytics information systems (e.g., Google Ads, CRM data, system logs, chat logs, call center logs, and recommender system data), and datasets of different levels of skewness. However, as user analytics metrics are often standard across analytics information systems, the process presented here can facilitate this research. Future work concerning the gold standard baseline also needs to be done. What is the ‘truth’ for personification for user representation is a research problem in itself due to the ‘curse of dimensionality – i.e., as more attributes are added, the representativeness usually decreases. It would also be an interesting extension to conduct a user study to explore how the transparency of the personification process affects decision makers’ selection and use of the personas.

There are also other impactful future research directions. The effect of hyperparameter selection on different attributes than those investigated here and ML models other than NMF should be explored by future work in the context of personification. Specifically, clustering needs to be investigated as it is commonly used for both user segmentation and personification (Salminen et al., 2021). In conjunction with this, metrics other than accuracy could be explored, such as diversity, novelty, fairness, or impact. As a further aspect, in this research, we only examined 5 to 15 hyperparameters. An exciting

avenue of research would be to push the upper limit to much higher numbers, as modern analytics information systems allow filtering and searching personas and other human-like representations of user data. Another area of inquiry could be investigating whether or not decision-makers using persona analytics systems understand the effects of hyperparameters tuning on the data in which they take action. However, for all these lines of investigation and research goals, the techniques employed in this research should be applicable.

## 7 Conclusions

This research explores the impact of hyperparameter selection on the accuracy of personifying user analytics data in a persona analytics system. Using tens of millions of user interactions from three industry-standard online channels for three organizations and employing a factorization ML model, we alter the hyperparameter of the number of personas in the set from five to fifteen. We compare age, nationality, and GAN to all personas from the resulting factorization baseline. The findings show that hyperparameter selection significantly alters the personification for all four user attributes, although the effect is most apparent with gender. The hyperparameters of five and ten personas provide an acceptable representation of the user population across all the attributes, implying these hyperparameters might be a good rule of thumb. These findings offer a foundation for future research in investigating the personification of user analytics data.

**Acknowledgements** We thank the corporations that provided access to the user analytics data employed in this research. The academic-industry partnership is a mutually beneficial and rewarding experience for all.

**Funding** Open Access funding provided by the Qatar National Library.

**Data Availability** The data used in this research is proprietary and can not be shared.

## Declarations

**Conflict of Interest** Authors declare no conflict of interest.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will



need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Agarwal, R., & Dhar, V. (2014). Big data, data science, and analytics: The opportunity and challenge for IS research. *Information Systems Research*, 25(3), 443–448.
- Agrawal, T. (2020). *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient* (1st ed. edition). Apress.
- Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, 1142(1), 012012. <https://doi.org/10.1088/1742-6596/1142/1/012012>.
- An, J., Kwak, H., Jung, S., Salminen, J., & Jansen, B. J. (2018a). Customer segmentation using online platforms: Isolating behavioral and demographic segments for persona creation via aggregated user data. *Social Network Analysis and Mining*, 8(1), 54. <https://doi.org/10.1007/s13278-018-0531-0>
- An, J., Kwak, H., Salminen, J., Jung, S., & Jansen, B. J. (2018b). Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Transactions on the Web (TWEB)*, 12(4), 27. <https://doi.org/10.1145/3265986>
- Arora, D., & Malik, P. (2015). Analytics: Key to Go from Generating Big Data to Deriving Business Value. *IEEE First International Conference on Big Data Computing Service and Applications, 2015*, 446–452. <https://doi.org/10.1109/BigDataService.2015.62>
- Bijmolt, T. H. A., Leeflang, P. S. H., Block, F., Eisenbeiss, M., Hardie, B. G. S., Lemmens, A., & Saffert, P. (2010). Analytics for Customer Engagement. *Journal of Service Research*, 13(3), 341–356. <https://doi.org/10.1177/1094670510375603>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan), 993–1022.
- Blomquist, A., & Arvola, M. (2002). Personas in action: Ethnography in an interaction design team. *Proceedings of the Second Nordic Conference on Human-Computer Interaction*, 197–200.
- Celebi, M. E., & Aydin, K. (Eds.). (2016). *Unsupervised Learning Algorithms* (1st ed. 2016 edition). Springer.
- Chang, Y., Lim, Y., & Stolterman, E. (2008). Personas: From Theory to Practices. *Proceedings of the 5th Nordic Conference on Human-Computer Interaction: Building Bridges*, 439–442. <https://doi.org/10.1145/1463160.1463214>.
- Chapman, C., Love, E., Milham, R. P., ElRif, P., & Alford, J. L. (2008). Quantitative Evaluation of Personas as Information. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 52, 1107–1111. <https://doi.org/10.1177/154193120805201602>
- Chapman, C., & Milham, R. P. (2006). The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50, 634–636. <https://doi.org/10.1177/154193120605000503>
- Chen, C., & Liu, L.-M. (1993). Joint estimation of model parameters and outlier effects in time series. *Journal of the American Statistical Association*, 88(421), 284–297.
- Chien, S.-Y., Lin, Y.-L., & Chang, B.-F. (2022). The Effects of Intimacy and Proactivity on Trust in Human-Humanoid Robot Interaction. *Information Systems Frontiers*, 1–16.
- Choi, B., Park, M., & Chai, S. (2016). Effect of Emotional Elements in Personal Relationships on Multiple Personas from the Perspective of Teenage SNS Users. *Information Systems Review*, 18(2), 199–223.
- Cohen, R. J. (2014). Brand Personification: Introduction and Overview. *Psychology & Marketing*, 31(1), 1–30. <https://doi.org/10.1002/mar.20671>
- 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.
- Darliansyah, A., Naeem, M. A., Mirza, F., & Pears, R. (2019). SENTI-PEDE: A Smart System for Sentiment-based Personality Detection from Short Texts. *Journal of Universal Computer Science*, 25(10), 1323–1352.
- Delbaere, M., McQuarrie, E. F., & Phillips, B. J. (2011). Personification in Advertising. *Journal of Advertising*, 40(1), 121–130. <https://doi.org/10.2753/JOA0091-3367400108>
- Denizci Guillet, B. (2020). Online upselling: Moving beyond offline upselling in the hotel industry. *International Journal of Hospitality Management*, 84, 102322. <https://doi.org/10.1016/j.ijhm.2019.102322>.
- Ditton, E., Swinbourne, A., Myers, T., & Scovell, M. (2021). Applying Semi-Automated Hyperparameter Tuning for Clustering Algorithms. ArXiv:2108.11053.
- Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. <https://doi.org/10.1145/3301275.3302310>.
- Drozdal, J., Weisz, J., Wang, D., Dass, G., Yao, B., Zhao, C., Muller, M., Ju, L., & Su, H. (2020). Trust in AutoML: Exploring information needs for establishing trust in automated machine learning systems. *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 297–307. <https://doi.org/10.1145/3377325.3377501>.
- Faily, S., & Flechais, I. (2011). Persona Cases: A Technique for Grounding Personas. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2267–2270. <https://doi.org/10.1145/1978942.1979274>.
- Fan, S., Lau, R. Y., & Zhao, J. L. (2015). Demystifying big data analytics for business intelligence through the lens of marketing mix. *Big Data Research*, 2(1), 28–32.
- Gil, Y., Honaker, J., Gupta, S., Ma, Y., D'Orazio, V., Garijo, D., Gadevar, S., Yang, Q., & Jahanshad, N. (2019). Towards human-guided machine learning. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 614–624. <https://doi.org/10.1145/3301275.3302324>.
- Griva, A., Bardaki, C., Pramatar, K., & Doukidis, G. (2021). Factors Affecting Customer Analytics: Evidence from Three Retail Cases. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-020-10098-1>
- Griva, A., Bardaki, C., Pramatar, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, 100, 1–16. <https://doi.org/10.1016/j.eswa.2018.01.029>
- Grudin, J., & Pruitt, J. (2002). Personas, Participatory Design and Product Development: An Infrastructure for Engagement. *Proceedings of Participation and Design Conference (PDC2002)*, 8.
- Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2125–2126. <https://doi.org/10.1145/2939672.2945386>.
- Helgesen, Ø. (2006). Customer segments based on customer account profitability. *Journal of Targeting, Measurement and Analysis for Marketing*, 14(3), 225–237. <https://doi.org/10.1057/palgrave.jt.5740183>

- Hertzberg, H. T., Daniels, G. S., & Churchill, E. (1954). *Anthropometry of flying personnel-1950*. Antioch Coll Yellow Springs OH.
- Holgersson, J., Alenljung, B., & Söderström, E. (2015). User participation at a discount: Exploring the use and reuse of personas in public service development. *European Conference on Information Systems (ECIS)*, paper-30.
- Hossain, M. A., Akter, S., & Yanamandram, V. (2020). Revisiting customer analytics capability for data-driven retailing. *Journal of Retailing and Consumer Services*, 56, 102187. <https://doi.org/10.1016/j.jretconser.2020.102187>.
- Huang, X., Wu, L., & Ye, Y. (2019). A Review on Dimensionality Reduction Techniques. *International Journal of Pattern Recognition and Artificial Intelligence*, 33(10), 1950017. <https://doi.org/10.1142/S0218001419500174>
- Ibnu, C. R. Muh., Santos, J., & Surendro, K. (2019). Determining the Neural Network Topology: A Review. *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, 357–362. <https://doi.org/10.1145/3316615.3316697>.
- Iivari, J., & Iivari, N. (2011). Varieties of user-centredness: An analysis of four systems development methods. *Information Systems Journal*, 21(2), 125–153.
- Iivari, N. (2009). “Constructing the users” in open source software development: An interpretive case study of user participation. *Information Technology & People*, 22(2), 132–156.
- Jansen, B. J., & Clarke, T. B. (2017). Conversion potential: A metric for evaluating search engine advertising performance. *Journal of Research in Interactive Marketing*, 11(2), 142–159. <https://doi.org/10.1108/JRIM-07-2016-0073>
- Jansen, B. J., Jung, S., Ramirez Robillos, D., & Salminen, J. (2021a). Too Few, Too Many, Just Right: Creating the Necessary Number of Segments for Large Online Customer Populations. *Electronic Commerce Research and Applications*, 101083. <https://doi.org/10.1016/j.elerap.2021.101083>.
- Jansen, B. J., Jung, S., & Salminen, J. (2019a). Capturing the change in topical interests of personas over time. *Proceedings of the Association for Information Science and Technology*, 56(1), 127–136.
- Jansen, B. J., Jung, S., & Salminen, J. (2019b). Creating Manageable Persona Sets from Large User Populations. *Extended Abstracts of the 2019b CHI Conference on Human Factors in Computing Systems*, 1–6. <https://doi.org/10.1145/3290607.3313006>.
- Jansen, B. J., Jung, S., & Salminen, J. (2020a). From flat file to interface: Synthesis of personas and analytics for enhanced user understanding. *Proceedings of the Association for Information Science and Technology*, 57(1). <https://doi.org/10.1002/pra2.215>.
- Jansen, B. J., Salminen, J., & Jung, S. (2020b). Data-Driven Personas for Enhanced User Understanding: Combining Empathy with Rationality for Better Insights to Analytics. *Data and Information Management*, 4(1), 1–17. <https://doi.org/10.2478/dim-2020-0005>
- Jansen, B. J., Sobel, K., & Cook, G. (2011). Classifying ecommerce information sharing behaviour by youths on social networking sites. *Journal of Information Science*. <https://doi.org/10.1177/0165551510396975>
- Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11), 2169–2188. <https://doi.org/10.1002/asi.21149>
- Jansen, B., Jung, S., & Salminen, J. (2021b). The Effect of Hyperparameter Selection on the Personification of Customer Population Data. *International Journal of Electrical and Computer Engineering Research*, 1(2), 2. <https://doi.org/10.53375/ijecer.2021.31>.
- Jansen, B., Salminen, J., Jung, S., & Guan, K. (2021c). *Data-Driven Personas* (1st ed., Vol. 14). Morgan & Claypool Publishers.
- Jung, S., An, J., Kwak, H., Ahmad, M., Nielsen, L., & Jansen, B. J. (2017). Persona Generation from Aggregated Social Media Data. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1748–1755.
- Jung, S., Salminen, J., & Jansen, B. J. (2021). All About the Name: Assigning Demographically Appropriate Names to Data-Driven Entities. *Proceedings of the 54th Hawaii International Conference on System Sciences*. <http://hdl.handle.net/10125/71108>.
- Jung, S., Salminen, J., Kwak, H., An, J., & Jansen, B. J. (2018). Automatic Persona Generation (APG): A Rationale and Demonstration. *CHIIR '18: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 321–324. <https://doi.org/10.1145/3176349.3176893>.
- Kalliola, J., Kapočiūtė-Dzikiene, J., & Damaševičius, R. (2021). Neural network hyperparameter optimization for prediction of real estate prices in Helsinki. *PeerJ Computer Science*, 7, e444. <https://doi.org/10.7717/peerj-cs.444>.
- Karumur, R. P., Nguyen, T. T., & Konstan, J. A. (2018). Personality, user preferences and behavior in recommender systems. *Information Systems Frontiers*, 20(6), 1241–1265.
- Kelleher, J. D., Namee, B. M., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies*. MIT Press.
- Kim, E., Yoon, J., Kwon, J., Liaw, T., & Agogino, A. M. (2019). From Innocent Irene to Parental Patrick: Framing User Characteristics and Personas to Design for Cybersecurity. *Proceedings of the Design Society: International Conference on Engineering Design*, 1(1), 1773–1782.
- Kitchens, B., Dobolyi, D., Li, J., & Abbasi, A. (2018). Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data. *Journal of Management Information Systems*, 35(2), 540–574. <https://doi.org/10.1080/07421222.2018.1451957>
- Laporte, L., Slegers, K., & De Grooff, D. (2012). Using Correspondence Analysis to Monitor the Persona Segmentation Process. *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, 265–274. <https://doi.org/10.1145/2399016.2399058>.
- Lauren Sorenson. (2011, December 13). *6 Core Benefits of Well-Defined Marketing Personas* Lauren Sorenson. <https://blog.hubspot.com/blog/tabid/6307/bid/29583/6-core-benefits-of-well-defined-marketing-personas.aspx>.
- Lee, D. D., & Seung, S. H. (1999). Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature*, 401(6755), 788–791.
- Lin, C., & Kunnathur, A. (2019). Strategic orientations, developmental culture, and big data capability. *Journal of Business Research*, 105, 49–60.
- Liu, H., Yao, L., Zheng, Q., Luo, M., Zhao, H., & Lyu, Y. (2020). Dual-stream generative adversarial networks for distributionally robust zero-shot learning. *Information Sciences*, 519, 407–422. <https://doi.org/10.1016/j.ins.2020.01.025>
- Maté, A., Trujillo, J., & Mylopoulos, J. (2017). Specification and derivation of key performance indicators for business analytics: A semantic approach. *Data & Knowledge Engineering*, 108, 30–49. <https://doi.org/10.1016/j.datak.2016.12.004>
- Meissner, F., & Blake, E. (2011). Understanding culturally distant end-users through intermediary-derived personas. *Proceedings of the South African Institute of Computer Scientists and Information Technologists Conference on Knowledge, Innovation and*

- Leadership in a Diverse, Multidisciplinary Environment - SAIC-SIT '11*, 314. <https://doi.org/10.1145/2072221.2072266>.
- Mijač, T., Jadrić, M., & Ćukušić, M. (2018). The potential and issues in data-driven development of web personas. *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1237–1242. <https://doi.org/10.23919/MIPRO.2018.8400224>.
- Mohamed, A. E. (2017). Comparative Study of Four Supervised Machine Learning Techniques for Classification. *International Journal of Applied Science and Technology*, 7(2), 14.
- Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436–465.
- Molenaar, L. (2017). *Data-driven personas: Generating consumer insights with the use of clustering analysis from big data*. Undefined. /paper/Data-driven-personas%3A-Generating-consumer-insights-Molenaar/d9c8d7adb6d4c1c2ab1f7c95c202c6770879c57b.
- Murray, P. W., Agard, B., & Barajas, M. A. (2017). Market segmentation through data mining: A method to extract behaviors from a noisy data set. *Computers & Industrial Engineering*, 109, 233–252. <https://doi.org/10.1016/j.cie.2017.04.017>
- Nasraoui, O., Cerwinske, J., Rojas, C., & Gonzalez, F. (2007). Performance of Recommendation Systems in Dynamic Streaming Environments. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (Vol. 1–0, pp. 569–574). Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611972771.63>.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175.
- Nielsen, L. (2004). *Engaging personas and narrative scenarios* [PhD Thesis]. Samfundslitteratur.
- Nielsen, L. (2019). *Personas—User Focused Design* (2nd ed. 2019 edition). Springer.
- Nielsen, L., Hansen, K. S., Stage, J., & Billestrup, J. (2015). A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations. *International Journal of Sociotechnology and Knowledge Development*, 7(1), 45–61. <https://doi.org/10.4018/ijskd.2015010104>
- Park, D., & Kang, J. (2022). Constructing Data-Driven Personas through an Analysis of Mobile Application Store Data. *Applied Sciences*, 12(6), 6. <https://doi.org/10.3390/app12062869>.
- Pelau, C., Dabija, D.-C., & Ene, I. (2021). What makes an AI device human-like? The role of interaction quality, empathy and perceived psychological anthropomorphic characteristics in the acceptance of artificial intelligence in the service industry. *Computers in Human Behavior*, 122, 106855. <https://doi.org/10.1016/j.chb.2021.106855>.
- Probst, P., Boulesteix, A.-L., & Bischl, B. (2009). Tunability: Importance of Hyperparameters of Machine Learning Algorithms. *Journal of Machine Learning Research*, 32.
- Pruitt, J., & Grudin, J. (2003). Personas: Practice and Theory. *Proceedings of the 2003 Conference on Designing for User Experiences*, 1–15. <https://doi.org/10.1145/997078.997089>.
- Ramsey, P. H., Hodges, J. L., & Popper Shaffer, J. (1993). Significance probabilities of the wilcoxon signed-rank test. *Journal of Nonparametric Statistics*, 2(2), 133–153. <https://doi.org/10.1080/10485259308832548>
- Reinartz, W. J., & Kumar, V. (2000). On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing. *Journal of Marketing*, 64(4), 17–35.
- Rust, R. T., & Huang, M.-H. (2014). The Service Revolution and the Transformation of Marketing Science. *Marketing Science*, 33(2), 206–221. <https://doi.org/10.1287/mksc.2013.0836>
- Salminen, J., Froneman, W., Jung, S., Chowdhury, S., & Jansen, B. J. (2020a). The Ethics of Data-Driven Personas. *Extended Abstracts of the 2020a CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–9. <https://doi.org/10.1145/3334480.3382790>.
- Salminen, J., Guan, K., Jung, S.-G., & Jansen, B. J. (2021). A Survey of 15 Years of Data-Driven Persona Development. *International Journal of Human–Computer Interaction*, 0(0), 1–24. <https://doi.org/10.1080/10447318.2021.1908670>.
- Salminen, J., Jung, S., Chowdhury, S. A., Sengün, S., & Jansen, B. J. (2020b). Personas and Analytics: A Comparative User Study of Efficiency and Effectiveness for a User Identification Task. *Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI'20)*. <https://doi.org/10.1145/3313831.3376770>
- Salminen, J., Jung, S., Nielsen, L., Şengün, S., & Jansen, B. J. (2022). How does varying the number of personas affect user perceptions and behavior? Challenging the ‘small personas’ hypothesis! *International Journal of Human–Computer Studies*, 168, 102915. <https://doi.org/10.1016/j.ijhcs.2022.102915>.
- Salminen, J., Kaate, I., Kamel, A. M. S., Jung, S., & Jansen, B. J. (2020c). How Does Personification Impact Ad Performance and Empathy? An Experiment with Online Advertising. *International Journal of Human–Computer Interaction*, 0(0), 1–15. <https://doi.org/10.1080/10447318.2020.1809246>.
- Salminen, J., Yoganathan, V., Corporan, J., Jansen, B. J., & Jung, S.-G. (2019). Machine learning approach to auto-tagging online content for content marketing efficiency: A comparative analysis between methods and content type. *Journal of Business Research*, 101, 203–217. <https://doi.org/10.1016/j.jbusres.2019.04.018>
- Sheth, J. N., Sisodia, R. S., & Sharma, A. (2000). The antecedents and consequences of customer-centric marketing. *Journal of the Academy of Marketing Science*, 28(1), 55–66. <https://doi.org/10.1177/0092070300281006>
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 553–572.
- Siegel, D. A. (2010). The Mystique of Numbers: Belief in Quantitative Approaches to Segmentation and Persona Development. *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 4721–4732. <https://doi.org/10.1145/1753846.1754221>.
- Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability* (pp. 15–18). Palgrave Macmillan. [https://doi.org/10.1007/978-1-349-20568-4\\_5](https://doi.org/10.1007/978-1-349-20568-4_5).
- Spiliotopoulos, D., Margaris, D., & Vassilakis, C. (2020). Data-Assisted Persona Construction Using Social Media Data. *Big Data and Cognitive Computing*, 4(3), 3. <https://doi.org/10.3390/bdcc4030021>.
- Stevenson, P. D., & Mattson, C. A. (2019). The Personification of Big Data. *Proceedings of the Design Society: International Conference on Engineering Design*, 1(1), 4019–4028. <https://doi.org/10.1017/dsi.2019.409>
- Subrahmaniyan, N., Higginbotham, D. J., & Bisantz, A. M. (2018). Using Personas to Support Augmentative Alternative Communication Device Design: A Validation and Evaluation Study. *International Journal of Human–Computer Interaction*, 34(1), 84–97.
- Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. *Machine Learning*, 65(1), 247–271. <https://doi.org/10.1007/s10994-006-9449-2>
- Terragni, S., & Fersini, E. (2021). An Empirical Analysis of Topic Models: Uncovering the Relationships between Hyperparameters, Document Length and Performance Measures. *Proceedings*



- of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), 1408–1416.
- Thirumuruganathan, S., Emadi, N. A., Jung, S., Salminen, J., Robillos, D. R., & Jansen, B. J. (2023). Will they take this offer? A machine learning price elasticity model for predicting upselling acceptance of premium airline seating. *Information & Management*, 60(3), 103759. <https://doi.org/10.1016/j.im.2023.103759>.
- Thirumuruganathan, S., Jung, S., Ramirez Robillos, D., Salminen, J., & Jansen, B. J. (2021). Forecasting the nearly unforecastable: Why aren't airline bookings adhering to the prediction algorithm? *Electronic Commerce Research*. <https://doi.org/10.1007/s10660-021-09457-0>
- Thirumuruganathan, S., Rahman, H., Abbar, S., & Das, G. (2014). Beyond itemsets: Mining frequent featuresets over structured items. *Proceedings of the VLDB Endowment*, 8(3), 257–268. <https://doi.org/10.14778/2735508.2735515>.
- Venkatesubramanian, S., & Hill, T. R. (2010). An empirical investigation into the effects of web search characteristics on decisions associated with impression formation. *Information Systems Frontiers*, 12(5), 579–593.
- Wang, C. (2022). Efficient customer segmentation in digital marketing using deep learning with swarm intelligence approach. *Information Processing & Management*, 59(6), 103085. <https://doi.org/10.1016/j.ipm.2022.103085>.
- Wechsler, J., & Schweitzer, J. (2019). Creating Customer-Centric Organizations: The Value of Design Artefacts. *The Design Journal*, 22(4), 505–527. <https://doi.org/10.1080/14606925.2019.1614811>
- Wedel, M., & Kannan, P. K. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, 80(6), 97–121.
- Wright, P., & McCarthy, J. (2008). Empathy and Experience in HCI. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 637–646. <https://doi.org/10.1145/1357054.1357156>.
- Wu, R.-S., & Chou, P.-H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. *Electronic Commerce Research and Applications*, 10(3), 331–341. <https://doi.org/10.1016/j.elerap.2010.11.002>
- Xu, Z., Frankwick, G. L., & Ramirez, E. (2016). Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective. *Journal of Business Research*, 69(5), 1562–1566.
- Yoon, J., Yang, K.-C., Jung, W.-S., & Ahn, Y.-Y. (2021). Persona2vec: A flexible multi-role representations learning framework for graphs. *PeerJ Computer Science*, 7, e439. <https://doi.org/10.7717/peerj-cs.439>.
- Yuan, X., Lee, J.-H., Kim, S.-J., & Kim, Y.-H. (2013). Toward a user-oriented recommendation system for real estate websites. *Information Systems*, 38(2), 231–243.
- Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, 58(4), 102555. <https://doi.org/10.1016/j.ipm.2021.102555>.
- Zehlike, M., Sühr, T., Baeza-Yates, R., Bonchi, F., Castillo, C., & Hajian, S. (2022). Fair Top-k Ranking with multiple protected groups. *Information Processing & Management*, 59(1), 102707. <https://doi.org/10.1016/j.ipm.2021.102707>.
- Zhang, M., Jansen, B. J., & Chowdhury, A. (2011). Business engagement on Twitter: A path analysis. *Electronic Markets*, 21(3), 161. <https://doi.org/10.1007/s12525-011-0065-z>
- Zheng, T., Zhang, Y., & Wang, Y. (2022). Dynamic guided metric representation learning for multi-view clustering. *PeerJ Computer Science*, 8, e922. <https://doi.org/10.7717/peerj-cs.922>.

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Bernard J. Jansen** is a Principal Scientist in the artificial intelligence center of the Qatar Computing Research Institute. He is a West Point graduate with a Ph.D. in computer science from Texas A&M University. Professor Jansen is editor-in-chief of the journal *Information Processing & Management* (Elsevier).

**Soon-gyo Jung** is a software engineer focused on data-driven/data-intensive systems in the artificial intelligence center at Qatar Computing Research Institute (QCRI), Doha, Qatar. He is currently strongly interested in computational social science, especially exploring its significant impact on society and how people communicate and share their culture with others. He received a B.E. degree in computer software from Kwangwoon University, Seoul, Korea in 2014, and an M.S. degree in electrical and computer engineering from Sungkyunkwan University, Suwon, Korea, in 2016.

**Joni Salminen** is Dr. a faculty member of the School of Marketing and Communication, University of Vaasa, Finland, and he is also affiliated with the Turku School of Economics at the University of Turku. Previously, he was a scientist at the Qatar Computing Research Institute. His research interests relate to personas, human–computer interaction, online hate, user and customer segmentation, and social media platforms.