The Effect of Hyperparameter Selection on the Personification of Customer Population Data

Bernard J. Jansen [1, *], Soon-gyo Jung [1], Joni Salminen [1, 2]
[1] Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
[2] Turku School of Economics, Turku University, Turku, Finland
Email: jjansen@acm.org, sjung@hbku.edu.qa, jsalminen@hbku.edu.qa
[*] *Corresponding author*

Abstract: **We explore the effects of hyperparameter selections on the personification accuracy of customer analytics data from a corporate YouTube channel with an audience in the hundreds of thousands and customer interactions in the tens of millions. Using non-negative matrix factorization, we generate personas sets from 5 to 15 using the customer analytics data, with the number of personas being the changing hyperparameter. We then compare the gender, age, nationality, and topical interests of the personas across each of the 11 persona sets using the average of the 110 generated personas as the baseline. This analysis shows that hyperparameter selection significantly alters the personification of the analytics data, with the effect most apparent with age representation. The set of 10 personas provides one of the most accurate representations across all attributes, indicating that this may be a good default hyperparameter for personification. Future research can explore other personification attributes with other customer analytics datasets.**

Keywords: **personification, customer data, analytics.**

I. INTRODUCTION

Many organizations desire to know more about their customers to make more informed business decisions. In pursuit of this goal, organizations often turn to customer analytics [1], [2]. The customer analytics process creates customer metrics that are proxies for real customers in customer populations that can now number in the millions to billions for companies. Businesses employ these customer analytics metrics for various purposes [3]–[6], such as marketing, advertising, product development, upselling, and customer relationship management. These customer analytic metrics are an example of *depersonification*, which we define in this context as "*the representation of real people by text, numerical, or other data creating one or more proxies for real people*".

With the increasing ease of collecting behaviors, purchases, the sentiment expressed in reviews and social media information, customer data volume has dramatically increased to a size that requires the employment of machine learning models to process and analyze [7]. Machine learning (ML) techniques are used with large volumes of data to identify trends and segments within big customer populations [8]. Robust persona analytics systems are now using these ML models to create data-driven personas, which are "humanized representation of people based on data about customers, audiences, or users". These data-driven personas are an example of *personification*, defined as "*the representation of numbers, metrics, and measures derived via analytics from customer data in the form of fictitious humans possessing attributes determined by the analytics process*".

The motivation for this personification [9] is that cold, rational numbers often do not generate the connection to and empathy of customers required [10] in endeavors such as marketing, advertising, product design, and service blueprinting. So, we are presented with the fascinating conundrum of organizations employing *depersonification* via customer analytics to make more informed decisions about these customers. These same organizations are then using *personification* via persona analytics to better relate to and empathize with these customers. This *depersonification-personification* concept is illustrated in Fig. 1.

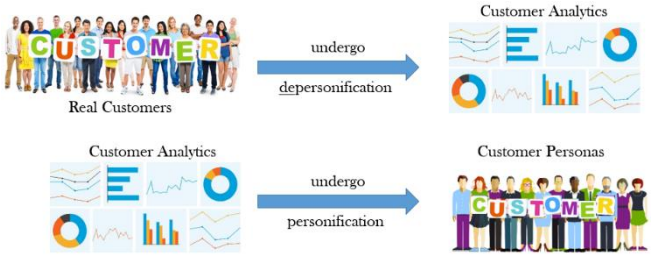

Fig. 1. Illustration of the on-going depersonification process to generate customer analytics motivated by a desire to make more informed business decisions about customers. While this depersonification process is occurring, a personification process is ongoing where customer analytics data is transformed into data-driven personas motivated by a desire to make empathic and understanding decisions about customers. Both depersonification and personification processes employ similar machine learning approaches for organizations with large customer populations.

The *depersonification* and *personification* processes typically employ ML approaches to generate customer analytics [11] or data-driven personas [12], [13] due to the large volume of customer data. There are many ML approaches, including supervised and unsupervised learning.

Nearly all ML models require the configuration of at least two parameters [14], model parameters and hyperparameters. A model parameter, usually just referred to as a parameter, is a configuration variable *internal* to the model where one can estimate or calculate the value from data (e.g., average, variance, maximum, minimum, mean). A hyperparameter, conversely, is a configuration variable *external* to the model where one cannot estimate the value from data (e.g., number of clusters, number of iterations). Hyperparameters are set manually, typically using heuristics or results from trial and

error. This setting procedure is because one usually does not know beforehand what the best value for a hyperparameter for a given model for a given problem for a given dataset is. Therefore, ML models [15] are often tuned to find the best or at least a reasonable value for a model's hyperparameters that result in the best performance.

Traditionally, personas were created manually based on limited customer data, requiring no ML approaches. However, persona analytics systems can theoretically create an arbitrary number of personas [16] from customer analytics representing thousands to millions to billions of customers. However, as an emerging field of research, the impact of hyperparameters on models employed by persona analytics has seen little research. In creating data-driven personas via persona analytics, the critical hyperparameter is the number of personas to create. Therefore, there are several unanswered questions, including: *What is the effect of changing the number of personas?*, *Does the accuracy of customer representation correlate with the 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.

This research uses a large customer analytics dataset numbering in the hundreds of thousands of customers and millions of interactions from a major international corporation's YouTube channel. We generate sets of personas from 5 to 15 from this customer analytics dataset using an identical ML approach. We then analyze the resulting 110 personas comparing each of the 11 groups of personas along four attributes to determine the most appropriate hyperparameter for creating data-driven personas.

We discuss theoretical and practical implications and present a literature review, research questions, methodology, and results. We end with a critique of this study and directions for future research.

II. REVIEW OF LITERATURE

This research is at the intersection of three generally independent research domains – machine learning, digital analytics, and personification.

As the literature of these fields is voluminous, we focus on the hyperparameter aspects of ML, the customer representation in the digital analytics domain, and data-driven personas in the personification field. Even with this narrowing, the prior literature presented here skims the surface, only presenting what is needed. However, the literature reviewed serves as a foundation for the research presented in this article. For deeper dives into each of these areas' specifics, we refer the interested reader to specific literature in each field.

A. Hyperparameters and Their Effect on ML Models

There are various approaches to ML [17], which is the use of computing algorithms that 'learn' (i.e., improve performance) by iterations using data. The approaches generally involve how much supervision the models are given or whether or not the specific ML models use statistical techniques [18], [19]. Each ML approach has many alternative models, which is a particular instance of that approach. For example, one approach to ML is factorization, and one particular factorization model is non-negative matrix factorization [20], which is the ML model used in this research. As customer data availability, accessibility, and volume have increased, ML models are increasingly employed to improve customer data analysis [21].

Most ML models have parameters that define the approach and the limits of the algorithmic analysis[22]. Parameters are usually derived by the models from the data that the model uses. However, these ML models also have hyperparameters [23], which also define the model's limits [24] but are not derived directly from the data. Hyperparameters are outside the data [25], usually derived from trial and error, estimation, or rules of thumb. For example, clustering is often used to segment customer analysis data [26], [27]. However, the 'right' number of clusters [26] is a matter of opinion or related to some data-external business goal. Therefore, most clustering algorithms require the number of clusters desired as a hyperparameter. Selecting the appropriate hyperparameter requires tuning.

This research generates data-driven personas from customer analytics data using a ML model as an example of personification. The number of data-driven personas to create from the data is a hyperparameter. In this research, we evaluate the effect of the hyperparameter values on the representation by the underlying customer analytics personification process.

B. Customer Analytics

Customer analytics data (when we say 'customer', we also mean 'user' or 'audience') is used for a variety of purposes [28] for the analysis and reporting concerning the consumers of a business's product or service [29]–[31]. Analytics 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 products sold) or ratios (e.g., the conversion rate for an ad which is the number of products sold due to the ad divided by the number of clicks on the ad). Hopefully, these numbers are evaluated as key performance indicators (KPIs)[32], and, in turn, businesses use these KPIs as measurements for progress toward achieving some business goal.

Companies employ customer analytics for various purposes [33]–[36], including branding, reputation management, customer relations management, marketing, advertising, cross-selling, upselling, and other business-related activities. If the data is at the individual level, one can employ it for personalization efforts, such as email marketing, for example. However, one often wants to aggregate the data to identify segments for tasks such as look-alike analysis or recommended products. Segmenting is the process of identifying groups of customers in the dataset that have common behaviors or demographic attributes [37].

Although it is beneficial and actionable for many tasks, customer analytics data is still a proxy for the real customers – that is, the actual customers have been *de-personified* to numbers. The reliance on customer analytics can lead to a

focus on the numbers at the expense of genuine customer focus, customer understanding, and customer empathy. Yet, empathy has been shown to provide beneficial effects for design and customer-centric decision making [38]. For this enhanced customer understanding, personification techniques and artifacts are advantageous.

For this research, we leverage customer analytics data to create one personification type, namely data-driven personas.

C. Data-Driven Personas

Personas depict fictional people created to represent real users, audiences, and/or customers [39]. Personas are an example of personification. Traditionally created somewhat *ad hoc* from limited data sources [40], the availability of large amounts of online customer data has opened the door for creating data-driven personas [41]. These data-driven personas are personas algorithmically created from actual customer data – usually large amounts of real customer data [42]. As such, data-driven personas are acknowledged to be precise and accurate representations via personification of the data. As data-driven personas are often created via persona analytics systems [43], these types of personas can be updated frequently – so they do not stale or become outdated. Therefore, data-driven personas provide humanized customer representations to make better decisions about customers and achieve better business results.

The concept of data-driven personas has been presented in-depth in groundbreaking books [44], [45] and later articles [13], [46], [47] on the subject and then extended by numerous researchers [48]–[50]. Data-driven personas have been investigated from several angles, including culturally [51] and as the representation of customer segmentation [13]. Data-driven personas have been shown to be more effective than straight-up analytics for certain customer-focused tasks [52]. Data-driven personas have also been shown to be quite effective in presenting an accurate customer representation [53] and effective in alternating the incorrect preconceptions of customers for stakeholders in organizations [54]. As such, data-driven personas are an excellent instantiation of the personification process. For more on data-driven personas, there are several published manuscripts on the topic [41], [43], [55].

This research leverages a ML model [20] and supporting algorithms [56]–[58] to generate data-driven personas [55] from an extensive customer analytics data set for personification. We specifically investigate the effect of different hyperparameter values of the number of personas on the personification process in representing the underlying data.

III. RESEARCH QUESTIONS

Our major research question (RQ) is: "Does the selection of hyperparameters affect the accuracy of personification of customer analytics data?". Specifically, we examine:

- RQa: *Does the selection of hyperparameters affect the accuracy of personification of customer analytics gender data?*
- RQb: *Does the selection of hyperparameters affect the accuracy of personification of customer analytics age data?*
- RQc: *Does the selection of hyperparameters affect the accuracy of personification of customer analytics nationality data?*
- RQd: *Does the selection of hyperparameters affect the accuracy of personification of customer analytics interest data?*

Our premise is that the hyperparameter values will affect the personification process, based on prior work in the ML area, combined with our experience in customer and persona analytics. Concerning accuracy, we mean *the degree to which the results of the hypermeter selection conform to the values from the baseline*.

We select the range of hyperparameter values from a floor of 5 to a ceiling of 15, as the choice represents a 3x increase/decrease in value, is the range of the persona analytics system that we employ in this research, and is within the cognitive capabilities of the stakeholders who will have to employ personas. It is also a range that persona studies commonly apply. Therefore, the hyperparameter values are realistic for actual employment. We select gender, age, and nationality for demographic varieties to investigate as these are industry standard attributes common in many industry-standard analytics platforms. We note that the major analytics platforms use biological sex as a proxy for gender and IP location for nationality.

Additionally, the demographic variables have different inherent values ranging from reasonably small with gender, limited with age (when condensed into age groupings), and relatively large with nationalities.

We choose topics of interest as it is derived directly from interaction with the products (i.e., videos), representing a customer behavioral attribute. In sum, our research questions represent a span of both demographic and behavioral characteristics to evaluate the effect of hyperparameter value selection on the personification process for accurately personifying customer analytics data.

IV. DATA COLLECTION AND METHODOLOGY

A. Data Collection

We employed the YouTube channel of a major international multi-billion US dollar aviation company for the customer analytics data. This company's YouTube channel has thousands of posted content pieces, more than 255,000 followers, with more than 84 million customer interactions (i.e., views) with content access worldwide. Therefore, ours is a robust, extensive, and heterogeneous customer analytics dataset.

As such, the company is representative of organizations with a large and diverse customer population. The customer data was accessed via YouTube Analytics, an industry-standard analytics platform. It is similar in design and metrics offered as other major analytics platforms, such as Google Analytics, Facebook Insights, Adobe Analytics, and IBM Analytics. The YouTube Analytics data is aggregated with

personal identifying attributes or values, so privacy concerns are minimal. However, the process employed in this research can be used with proprietary customer relationship management (CRM) data, which does contain individual customer information.

Also, the YouTube Analytics data is available only to the channel account holder and not open to the general public. The data for this research is employed with the channel account holder's permission, who provided access to the data via an Application Programming Interface (API) to the analytics platform.

B. Personification Process

We employ an industry-standard persona analytics system for automatic persona generation (APG) for the personification process, available online at https://persona.qcri.org. APG takes large amounts of customer analytics data and personifies the data via the creation of data-driven personas that are, reportedly, accurate and precise representations of the underlying customer data. As the persona analytics system [42], [59]–[61] has been described in detail in other research, we briefly present it here and refer the interested reader to published research.

APG leverages customer analytics data from various possible sources, including CRM, Facebook, Instagram, Facebook Ads, Google Ads, Google Analytics, and YouTube, via accessing the APIs after inclusion in an organizational account of the particular platform. The customer analytics data accessed by APG is aggregated demographics data (e.g., gender, age group, country code) and behavioral data (e.g., product id, count of customer 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 for this research, the product is videos. The behavioral metric is the viewcount of a given video.

APG uses non-negative matrix factorization (NMF) [20] algorithm as the initial step in the personification of customer analytics data. NMF is a factorization approach to identify sets of products related to customer behaviors (e.g., videos typically viewed by the same types of customers). Via matrix decomposition, NMF builds a matrix of these products and latent features. NMF also makes a matrix of demographic groups and their association to some latent factor. NMF associates the demographic groups to the product sets using the latent factors resulting in textual and numerical user profiles. APG infers the gender, age, and nationality attributes from the data in the customer analytics dataset via the analytics platform.

APG then enriches these user profiles to generate complete and rich data-driven personas. APG employs an internal database of thousands of stock photographs of models, with each image meta tagged with an appropriate gender-age-nationality. The system also has an internal database 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 [62]. Leveraging 2nd and 3rd party data, such as from the Facebook Audience Manager, APG calculates the probability of various other background information, such as occupation, education, and relationship status. For the topics of interest, APG generates these topics of interest [63] based on the products interacted with to generate using a variety of classification algorithms, including zero-shot classification and supervised ML.

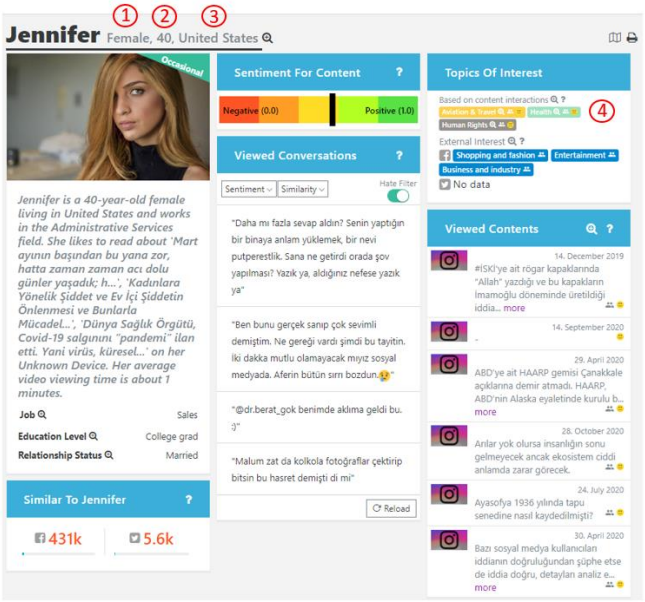

Fig. 2. Example of an APG data-driven persona profile (not from the organization used in this study) with the four elements of the persona profile pertinent to this study annotated. 1 – the persona gender. 2 – the persona age. 3 – the persona country. 4 – the persona topics of interest based on products or content from the specific organizational account.

C. Outcome of the Personification Process

Using various algorithmic approaches and periodic data collection, APG determines the sentiment of social media comments in multiple languages, including English, Arabic, Turkish, Spanish, Finnish, and French. An example of a data-driven persona created by the APG persona analytics systems is shown in Fig. 2, with the key features employed in this research annotated.

As a persona analytics system, APG has an advantage relative to manual approaches to persona creation. APG can generate a different number of personas in given persona sets or casts. Although theoretically, APG can create any number of personas, given the users' cognitive limitation of managing a large number of personas, the interface affords generation of persona sets from 5 through 15 inclusive. The sets of personas are presented as a scrolling window on the left of the interface. The set's personas are default ranked by the size of the customer segment they represent, with an example of 5 personas shown in Fig. 3.

Fig. 3. Example of the APG interface, with the cast of personas (see 1 – in this case, a set of 5 personas). The personas in the listing are ranked by the current reach of the personas (see 2), which is the current customer segment's size. Clicking on a result in the persona listing (see 1, the shaded result at row one) displays the corresponding persona profile (see 3, in this case, Abhishek). Note: The grayed areas in the persona profile are instances of the company name or identifying information that has been masked to preserve the company's confidentiality.

D. Changing of Hyperpapameter Values

As the number of personas is increased from 5 to 15, the persona listing increases correspondingly, with the underlying customer population being personified with greater granularity (see Table I). As shown in Table I, the hyperparameter (i.e., the number of personas generated) is increased from the five personas shown in Fig. 3 to six and above.

For each hyperparameter selection, APG generates a new listing of data-driven personas. The default ordering of the personas is the size of the customer segment that each persona represents. As the generation of personas from 5 to 15 is based on the *same* underlying customer population data, each of the persona listings are 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 (i.e., the set of 5 personas is based on the same customer interaction data as the set of 15 personas, for example). However, the same or similar personas can appear in multiple persona listings when the hyperparameter values change.

From reviewing the casts of personas in Fig 3 and Table I, we see that some portions of the persona listings intersect (i.e., the identical/similar personas are in multiple listing), and other personas are unique (i.e., the personas only appear in one in one persona listing). Also, as the persona analytics system segments the customer analytics data based on the hyperparameter values, the ranking of some of the personas that may represent similar segments across listings changes.

This ranking change results from the changing size and number of the customer segment represented by each of the personas in each set. We also note that some personas and their underlying customer 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, such as Abhihek (29 years old from India) and Indra (30 years old from Indonesia). Other personas appear in multiple listings but at different rankings, such as Chris (29 years old from the United States). In sum, a visual inspection of the persona listings indicates that the hyperparameters' change might skew user perceptions of the customer population.

However, whether five, fifteen, or any number of personas in between, the total of the personas in the listing represent the total of the underlying customer analytics dataset. Therefore, each representation is, in a sense, correct. However, in order to compare across sets of personas, we need a baseline. We generate all 11 persona sets for this baseline and then calculate the proportional representation for gender, age, nationality, and topics of interest. This gives us a 'gold standard' baseline that we take the 'accurate' representation. We then calculate the expected number of personas that we would expect in each set with these percentages. We compared the *expected* number to the *actual* number that was generated in each set. We then determine if the actual number of personas exceeds (Over), is less than (Under), or is correct (Same). We can then determine which persona sets are more accurate, and if not exact, whether they over or under-represented the customer population for the given attributes. One can calculate the times a given protected attribute is over or under, and this way, quantify "algorithmic bias" in data-driven personas.

We are not arguing that this averaging approach is the *best* for determining the accuracy of personification. This is not the objective of our research, and this determination of the 'best', if such a thing exists, we leave for other research. However, this baseline does suffice for our research to determine if changing the hyperparameter' values affects the personification process. As such, the baseline is sufficient for investigating our research questions.

TABLE I. CASTS OF DATA-DRIVEN PERSONAS FROM APG, A PERSONA ANALYTICS SYSTEM EMPLOYING NON-NEGATIVE MATRIX FACTORIZATION TO CREATE PERSONAS SETS. SHOWN IN THIS TABLE ARE THE CASTS OF PERSONAS FROM 6 TO 15, WITH THE CAST OF 5 PERSONAS SHOWN IN FIG. 3. THE PERSONAS IN EACH CAST ARE RANKED BY THE SIZE OF THE CUSTOMER SEGMENT EACH PERSONA REPRESENTS, FROM LARGEST TO SMALLEST SEGMENT. EACH PERSONA RESULT IN THE LISTINGS DISPLAYS NAME, AGE SELECTED FROM AGE CATEGORY, AND NATIONALITY, ALONG WITH A PICTURE. THE PICTURES AND NAMES ARE META-TAGGED WITH THE APPROPRIATE GENDER, AGE, AND NATIONALITY TO HELP ENSURE THE PERSONAS ARE BELIEVABLE.

Value of the Hyperparameter (i.e., number of data-driven personas generated from customer analytics data by APG) from 6 to 15, with the corresponding persona listing for each)				
6	**7**	**8**	**9**	**10**
Abhishek 29, India	Abhishek 29, India	Abhishek 29, India	Abhishek 29, India	Abhishek 29, India
Indra 30, Indonesia	Indra 30, Indonesia	Indra 30, Indonesia	Indra 30, Indonesia	Indra 30, Indonesia
Issam 32, Qatar	Issam 32, Qatar	Issam 32, Qatar	Issam 32, Qatar	Issam 32, Qatar
John 30, United States	Chris 39, United States	Chris 39, United States	Chris 39, United States	Chris 39, United States
Abdulrahman 28, Jordan	Abdulrahman 28, Jordan	Abdulrahman 28, Jordan	Abdulrahman 28, Jordan	Abdulrahman 28, Jordan
Budi 43, Indonesia	Budi 43, Indonesia	Budi 43, Indonesia	Budi 43, Indonesia	Budi 43, Indonesia
	Rahul 20, India	Rahul 20, India	Rahul 20, India	Rahul 20, India
		Jack 19, United Kingdom	Jack 19, United Kingdom	Jack 19, United Kingdor
			Amit 37, India	Amit 37, India
				John 30, United States
11	**12**	**13**	**14**	**15**
Abhishek 29, India	Abhishek 29, India	Abhishek 29, India	Abhishek 29, India	Abhishek 29, India
Indra 30, Indonesia	Indra 30, Indonesia	Indra 30, Indonesia	Indra 30, Indonesia	Indra 30, Indonesia
Issam 32, Qatar	Issam 32, Qatar	Rahul 20, India	Priya 26, India	Priya 26, India
Chris 39, United States	Rahul 20, India	Arm 24, Thailand	Chris 39, United States	Arm 24, Thailand
Abdulrahman 28, Jordan	Abdulrahman 28, Jordan	Abdulrahman 28, Jordan	Abdulrahman 28, Jordan	Abdulrahman 28, Jordan
Budi 43, Indonesia	Budi 43, Indonesia	Budi 43, Indonesia	Budi 43, Indonesia	Budi 43, Indonesia
Rahul 20, India	Shreya 24, India	Shreya 24, India	Rahul 20, India	Shreya 24, India
Jack 19, United Kingdom	Jack 19, United Kingdom	Jack 19, United Kingdom	Arm 24, Thailand	Lucas 26, Brazil
Amit 37, India	Amit 37, India	Amit 37, India	Shreya 24, India	Mostafa 21, Iraq
Arm 24, Thailand	Haziq 32, Malaysia	Haziq 32, Malaysia	Lucas 26, Brazil	Haziq 32, Malaysia
John 30, United States	Chris 39, United States	Chris 39, United States	Jack 19, United Kingdom	Jack 19, United Kingdom
	John 30, United States	John 30, United States	John 30, United States	John 30, United States
		Issam 32, Qatar	Issam 32, Qatar	Issam 32, Qatar
			Haziq 32, Malaysia	Usman 22, Pakistan
				Chris 39, United States

V. RESULTS

Here, we present the analysis results of varying the hyperparameter of the personas analytics system from sets of 5 to 15 personas on the accuracy outcomes of gender, nationality, and topics of interest. We begin with results from an exploratory analysis of the personas generated from the customer population dataset

A. Exploratory Analysis

We analyzed the 110 personas generated from the 11 generations of the persona sets from 5 to 15 personas.

There were 104 male personas (94.5%) and six female personas (5.5%). All analytics platforms 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 platforms. So, we reserve analysis for other gender identities for future research.

Concerning age, the average for the 110 personas was 29.95 years (sd = 6.80, max = 43, min = 19, mdn = 30), reflecting the younger customer population of the YouTube platform. For the analysis, we clustered the 110 personas in age groupings to reflect the age clusters in the underlying analytics platform, generally mirroring the US Census categories, resulting in three age groupings for the 110 personas. These three age groupings are: 18-24 age category (23.6% of the personas), 25-34 age category (54.5%), and 35-44 age category (21.8%).

The 110 personas are from eleven countries, with India (27.3%), Indonesia (19.1%), and the United States (15.5%) being the most common, followed by Jordan (10.0%), Qatar (10.0%), United Kingdom (7.3%), Malaysia (3.6%), Thailand (3.6%), Brazil (1.8%), Iraq (0.9%), and Pakistan (0.9%).

Concerning topics of interest, a single persona could have more than one topic of interest. There were 296 topical interests in total for the 110 personas, with Sports (26.7%) being the most popular. Other topics of interests in descending order are Business (13.2%), News (11.8%), Togetherness (10.8%), Community (10.1%), Cabin crew (9.5%), Contests (7.4%), Food (7.4%), and Airports (3.0%).

We now move to the analysis addressing our research questions concerning gender, age, nationality, and interests.

B. Gender

Concerning RQa (*Does the selection of hyperparameters affect the accuracy of personification of customer analytics gender data?*), as shown in Table II, most personas sets accurately represented genders but not all. Only three (20.0%) of the sets over/underestimate the gender distribution of the overall customer population, and eight (80.0%) of the sets presenting accurate representations. This finding is most likely the result of the heavily biased male representation. Sets 12 and 13 presented the most accurate representations that included females and not just male personas. Addressing our RQa, the number of personas does affect the accuracy of the personification of customer analytics gender data.

TABLE II. RESULTS OF GENDER ANALYSIS OF THE 11 PERSONA SETS OF 110 PERSONAS. BOLDED SETS (SETS 12 AND 13) ARE THE MOST ACCURATE REPRESENTATION OF THE CUSTOMER POPULATION. THE SHADED SETS (SETS 11, 14, AND 15) ARE THE LEAST ACCURATE REPRESENTATION OF THE CUSTOMER POPULATION. AS SHOWN, MOST PERSONA SETS WERE ACCURATE REPRESENTING GENDER OF THE CUSTOMER POPULATION

Gender	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Female	S	S	S	S	S	S	U	**S**	**S**	O	O
Male	S	S	S	S	S	S	O	**S**	**S**	U	U

TABLE III. RESULTS OF AGE ANALYSIS OF THE 11 PERSONA SETS OF 110 PERSONAS. BOLDED SETS (SETS 8, 10, AND 12) ARE THE MOST ACCURATE REPRESENTATION OF THE CUSTOMER POPULATION. THE SHADED SET (SET 5) IS THE LEAST ACCURATE REPRESENTATION OF THE CUSTOMER POPULATION.

Age	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
18-24	U	U	U	**S**	S	**S**	O	**S**	O	O	O
25-34	O	O	S	**S**	U	**S**	U	**S**	U	S	S
35-44	U	S	O	**S**	O	**S**	S	**S**	S	U	U

C. Age

Concerning RQb (*Does the selection of hyperparameters affect the accuracy of personification of customer analytics age data?*), as shown in Table III, the sets of 8, 10, and 12 personas provided the most accurate overall representation, with all three age groups accurate represented. The set of 5 personas was the least accurate, with no age groupings accurately represented. Addressing our RQb, the hyperparameter of the number of personas does affect the accuracy of the personification of customer analytics age data.

D. Country

As shown in Table IV, the sets of 11 and 12 personas provided the most accuracy for overall representation of the underlying data, while still including one or more of the smaller customer segments. Overall, all the persona sets were okay, with the lowest accuracy being 27.3% from the set of 15 personas. All other personas sets had accuracies of either 81.8% or 100%. One attribute of the underlying customer population is that nationalities are skewed toward a handful of countries with large representations in the customer population and many other nationalities with little representation. Given that there are eleven countries, one could perhaps expect the set of 5 personas not accurately to represent the customer population, but 15 was surprising. Possibly, due to the eleven countries, the proportional representation was skewed with this higher persona set over representing the smaller nationalities and under-representing the larger countries. This addresses RQc. The selection of hyperparameters does affect the accuracy of personification of customer analytics nationality data.

TABLE IV. RESULTS OF COUNTRY ANALYSIS OF THE 11 PERSONA SETS OF 110 PERSONAS. BOLDED SETS (SETS 12 AND 13) ARE THE MOST ACCURATE REPRESENTATIVE OF THE CUSTOMER POPULATION, WITH SOME OF THE SMALLER NATIONALITIES REPRESENTED. THE SHADED SET (SET 15) IS THE LEAST ACCURATE REPRESENTATION OF THE CUSTOMER POPULATION.

Country	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
India	S	U	S	S	S	S	S	**O**	**S**	S	U
Indonesia	S	O	O	S	S	S	S	**S**	**U**	U	U
United States	S	S	S	S	S	S	S	**S**	**S**	S	S
Jordan	S	S	S	S	S	S	S	**S**	**S**	S	U
Qatar	S	S	S	S	S	S	S	**S**	**S**	S	U
United Kingdom	S	S	U	S	S	S	S	**S**	**S**	S	S
Malaysia	S	S	S	S	S	S	U	**S**	**S**	S	S
Thailand	S	S	S	S	S	S	O	**U**	**O**	S	O
Brazil	S	S	S	S	S	S	S	**S**	**S**	O	O
Iraq	S	S	S	S	S	S	S	**S**	**S**	S	O
Pakistan	S	S	S	S	S	S	S	**S**	**S**	S	O

E. Interests

Examining topics of interest for RQd (*Does the selection of hyperparameters affect the accuracy of personification of customer analytics interest data?*), Table V shows that the presentation of customer population interests does change based on the hyperparameter setting. The sets of 9 and 13 personas are the least accurate representation, with persona sets 10 and 11 being the most accurate. Addressing our RQd, the hyperparameter of the number of personas does affect the accuracy of the personification of customer analytics for topics of interest.

TABLE V. RESULTS OF INTEREST ANALYSIS OF THE 11 PERSONA SETS OF 110 PERSONAS. BOLDED SET (SETS10 AND 11) ARE THE MOST ACCURATE REPRESENTATION OF THE CUSTOMER POPULATION. THE SHADED SETS (SETS 9 AND 13) ARE THE LEAST ACCURATE REPRESENTATION OF THE CUSTOMER POPULATION.

Topic	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Sports	O	O	O	O	S	**U**	**S**	U	U	S	U
Business	S	S	O	S	U	**S**	**S**	S	U	S	U
News	U	S	U	S	U	**S**	**O**	O	S	S	S
Together	U	S	U	S	U	**S**	**S**	S	S	O	S
Community	S	S	U	U	S	**S**	**S**	S	O	S	S
CabinCrew	S	S	S	O	O	**O**	**U**	S	S	U	S
Contest	U	U	S	S	O	**S**	**S**	S	O	S	O
Food	S	U	S	S	O	**S**	**S**	O	O	O	O
Airport	S	S	S	U	S	**S**	**S**	S	O	S	O

VI. DISCUSSION AND IMPLICATIONS

Our investigation examines the effect of changing the hyperparameter values of a persona analytics system through 11 iterations of values from 5 to 15 personas inclusive. Results show that changing the number of personas alters the personification representation of the customer analytics data along four attributes of gender, age, nationality, and topics of interest.

A. Personification by Gender

Concerning gender, there was a predominance of males (94.5%), limiting the hyperparameters' effect in the personification process – most of the personas were males regardless of the hyperparameters. For 110 personas in the 11 complete personas sets, the sets were correct representations 76.7% of the time and incorrect (either Under/Over-representation) 27.3% of the time. However, it was not until the set of 12 personas that a female persona appeared, indicating that if one wants to represent these smaller customer segments, one should generate more than a handful of personas.

B. Personification by Age

Concerning age grouping, the 25-34 age category represented about half (54.5%) of the baseline population. However, only three of the hyperparameters (8, 10, and 12 personas) resulted in accurate personification representations. The hyperparameter selection effect on age was quite pronounced, with 72.7% of the persona sets having an accuracy of 33% or less. The set of 5 personas over or under represented all age groups for 0.0% accuracy. For each entire 110 personas, age groups were accurately represented 48.5% of the time and over-represented 51.5% of the time.

C. Personification by Nationality

Concerning nationality, the changing of the hyperparameters also altered the personification representation of the customer population. However, it was not as drastic as with age, with most hyperparameter settings achieving 81.8% or above accurate representations, with 27.3% being the lowest (set 15). Nearly all personas sets were correct with the most represented countries, India, Indonesia, the United States, and Jordan. Most personas sets presented accurate representations of the other primary nationality, such as Qatar and the United Kingdom. However, the persona sets were less accurate with Malaysia, Thailand, Brazil, Iraq, and Pakistan, either over or under-representing them. Again, with skewed populations, one might lean toward a larger number of personas to personify the smaller segments. Overall, of the 110 personas, the representations were accurate 83.5% of the time. However, this 16.5% percent inaccuracy is impactful when considering the goal of depicting the diversity of the overall audience.

D. Personification by Interests

Moving to topical interests, again, the changing the hyperparameters of the personification of the customer analytics data resulted in changing the representation accuracy of the customer population. The sets of 9 and 13 personas had the least accurate representation (33.3%), with personas sets of 10 and 11 having the most accurate representation (77.8%). In general, the persona set hyperparameters equally over-represented (22.2%) and under-represented (21.2%) of the topical interests of the customer populations. However, it was not until the set of nine personas that all nine topics of interest were represented at least once. Again, this points to more personas rather than less for larger customer populations.

E. Verification with Second Customer Analytics Dataset

We were interested if our findings on the effect of hyperparameters on personification hold across customer datasets. Therefore, we repeated our analysis on a customer dataset from the Facebook page of the same organization, which has hundreds of thousands of followers and tens of

millions of interactions.

Using the identical methodology as previously, we again generated and analyzed the 110 personas generated from the 11 generations of the persona sets. There were 108 male personas (98.2%) and 2 female personas (1.8%). Concerning age, the average for the 110 personas was 28.48 (sd = 5.27, max = 42, min = 18, mdn = 30), again clustering the 110 personas in three age groupings of 18-24 age category (19.1%), 25-34 age category 78.2%, and 35-44 age category (2.7%). The 110 personas were from nine countries, with Bangladesh (19.1%) and the Philippines (19.1%) being the most common, followed by India (12.7%), Qatar (11.8%), Nepal (10.9%), Iraq (10.0%), Pakistan (10.0%), Cambodia (3.6%), and Saudi Arabia (2.7%). There were 386 topical interests for the 110 personas, with Sports (22.0%) being the most popular concerning topics of interest. Other topics of interests in descending order are Togetherness (12.7%), Business (10.9%), Contests (10.9%), Airports (10.1%), Food (10.1%), News (10.1%), Cabin crew (7.3%), and Community (6.0%). We again analyzed for gender, age, nationality, and interests.

The specific outcome of hyperparameter values varied (i.e., which persona sets were more or less accurate), as expected, with the second dataset. However, there were similar trends between the two datasets. The setting of the hyperparameters had the most notable effect on age again, with the persona sets in the middle (sets 9 and 10) being the most representative. The personification in terms of nationality was not as accurate, on average, as with our YouTube dataset, but it was still reasonable (55.6%). Interestingly, persona set 15 was again one of the least accurate representations of nationality. Again, the persona sets 10 and 12 were the most accurate in terms of conforming to the baseline for topics of interest.

Based on the comparison of this additional dataset, it confirms that tuning the value of hyperparameters for personification affects the accuracy of the representation of the customer data. It is also apparent that, for precise representation during personification, *the hyperparameters should be tuned to the specific dataset*, although there are some general trends, such as one should opt for a higher but not too high number of personas for both good accuracy and broader representation, shown in Fig. 4.

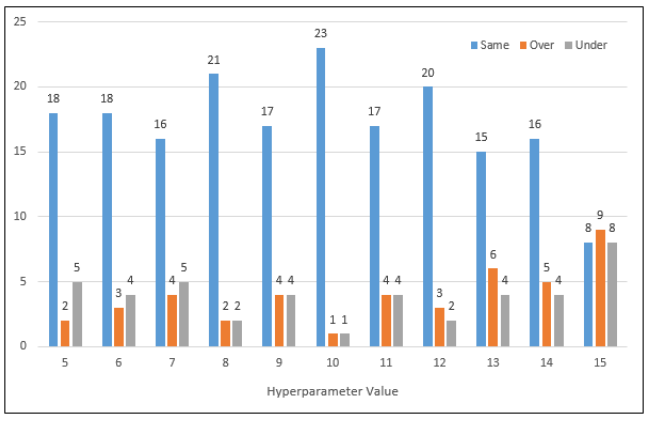

Fig. 4. The effect of hyperparameter selection on the total of the 25 data-driven person attributes, as measured by Same (i.e., identical with baseline), Over (i.e., over the baseline), and Under (i.e., under the baseline). As shown, the hyperparameter value of 10 is overall the most accurate. Interestingly, the 15 persona set, which one would expect to be the most granular, has the least accuracy. However, if a metric such as diversity was used, the larger hyperparameters might perform better.

F. Theoretical Implications

Concerning our core RQ (*Does the selection of hyperparameters affect the accuracy of personification of customer analytics data?*), the answer is definitely 'yes'. Of the four persona attributes studied, the hyperparameter selection during the personification process affected the accuracy of the customer population representation to varying degrees in all four cases.

This indicates that, as with other ML approaches, personification approaches employing big data from customer analytics to create human-like representations, such as personas, need to expend the effort to tune the ML model to the appropriate hyperparameters. Through trial and error, this tuning will enable the optimal hyperparameter for determining the number of personas that provides the most accurate representation of the underlying data.

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 customer population [64]. In this respect, our premise did not entirely hold. Generally, increasing the size of the personas set did commonly increase accuracy – up to a point. Then the accuracy of representation diminished, with the set of 15 personas often being among the least accurate representations. It has been noted in prior work [65], [66] that increased attributes decrease the representative [67]; however, given that the personas are actually created for the same analytics set, we still found this outcome surprising.

A portion of this could be the small number of categories for gender and age, along with the selection of the overall average for the baseline. This "flaw of averages" is documented in a classic study in 1950 conducted by US Air Force, finding that, among 4,000 measured pilots, no pilots matched all the average attributes of height, weight, etc. [68]. However, our findings imply a curvilinear function between hyperparameters for personalization and representation accuracy [69]. The determination of this function, we leave to future research.

G. Practical Implications

Concerning practical implications resulting from this research, we discuss the following three.

Avoiding the Mystique of Numbers: First, when employing personification techniques such as data-driven personas, stakeholders of personas need to be aware that a set of personas may give a biased view of the customer segmenting [70], [71] of the population, at least in terms of gender, age, nationality, and interests, as shown in this research. Just because the 'answer' involves a lot of data and an algorithm, it does not mean the 'answer' is 'accurate'. This ideal requires that developers of data-driven personas adequately execute tuning of the ML models. End users of the persona analytics systems must investigate whether or not the model hyperparameters were tuned before making marketing, advertising, content, or design decisions based on sets of personas derived from customer analytics data.

Ten is not Perfect, but Ten is Okay: Second, it appears that for the four attributes explored in this study, that *a set of 10 personas* may be a reasonable number as a 'rule of thumb'

for a generally accurate representation [72], [73] of the foundational customer populations data. In our analysis, the set of ten personas was either the or one of the most accurate personification representations (e.g., age and interests), so in the absence of tuning or initial hyperparameter selection, ten may be a good place to start. However, model tuning is still preferred to determine the hyperparameters in a data-driven manner.

Personification Needs an 'X': Third, this is not to say that ten is always the best, however. In this research, we were interested in an accurate representation. Using our baseline of the entire customer population's average, ten was generally accurate in presenting the customer population. However, they may be other business goals than accuracy, where the representation of the customer population average is not appropriate. For example, an organization may want to emphasize the diversity within a customer population [74] to highlight small segments for targeting or emerging segments. Businesses may wish to represent the most impactful segments [75] or the least costly segments [76]. For cases such as these, a more extensive set of personas would seem more appropriate. These numerical outliers are consumed or hidden within the 'average' of the overall customer population characteristics. However, this premise needs to be further investigated.

H.Strengths, Limitations, and Future Work

As with most research, there are both strengths and limitations. For strengths, we employed a large customer analytics dataset numbering in the tens of millions of customer actions from a large international corporation. So, the dataset is representative of those entities with extensive and diverse customer populations, implying that the overall approach is generalizable, at least within the frame of online audiences. The personification process of this customer data was accomplished using a state-of-the-art persona analytics system employing a factorization approach for personification of the analytics data. The process was validated with a second dataset. So, there are several strengths of the research.

Concerning limitations, we leveraged only one personification approach, data-driven personas. Other forms of personification, such as user profiles, should be examined to see how ML models in these approaches 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 approaches. Future work would need to confirm this, however.

For other limitations, we employed two data sets, one ML model, one metric (i.e., accuracy), four attributes (i.e., gender, age, nationality, interests), and one baseline (i.e., an average of the personification sets). Future work should explore the setting of personification hyperparameters on other customer data sets from other companies, other types (e.g., retail data or customer relations management data), and other platforms (e.g., Google Ads, Google Analytics, Facebook Insights). However, as customer analytics metrics are often common across platforms, 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 representation is a research problem in itself due to the 'curse of dimensionality, (i.e., as more attributes are added, the representativeness decreases).

There are many other future research directions. The effect of hyperparameter selection on different attributes [77] than the ones here and different ML models other than NMF should also be explored by future work in the context of this domain. In conjunction, the exploration of metrics other than accuracy could be explored, such as diversity, novelty, fairness, or impact. In this research, we examined hyperparameters from five to fifteen only. Interesting research would be to push the upper limit to much higher numbers, as modern personification analytics systems allow filtering and searching personas and other human-like representations of data. For all of these lines of investigation, however, the techniques employed in this research should be generalizable with these other business goals.

VII. CONCLUSION

This research explores the impact of hyperparameter selection on the accuracy of personifying customer analytics data. Using tens of millions of customer interactions from two social media channels 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 topical interests using the average of all of the personas as the standard baseline from the resulting factorization. Findings show that hyperparameter selection significantly alters the personification for all four factors, although the effect is most apparent with age. The hyperparameter of ten personas provides an acceptable representation of the customer population across all the attributes, implying ten might be a good rule of thumb. These findings offer a foundation for much future research in investigating the personification of customer analytics data.

ACKNOWLEDGMENT

We thank the corporation that provided access to the customer analytics data employed in this research. The academic-industry partner is a mutually beneficial rewarding experience for all.

REFERENCES

[1] S. Fan, R. Y. Lau, and J. L. Zhao, "Demystifying big data analytics for business intelligence through the lens of marketing mix," *Big Data Research*, vol. 2, no. 1, pp. 28–32, 2015.

[2] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2169–2188, 2009, doi: https://doi.org/10.1002/asi.21149.

[3] M. Wedel and P. K. Kannan, "Marketing analytics for data-rich environments," *Journal of Marketing*, vol. 80, no. 6, pp. 97–121, 2016.

[4] Z. Xu, G. L. Frankwick, and E. Ramirez, "Effects of big data analytics and traditional marketing analytics on new product success: A knowledge fusion perspective," *Journal of Business Research*, vol. 69, no. 5, pp. 1562–1566, 2016.

[5] M. Zhang, B. J. Jansen, and A. Chowdhury, "Business engagement on Twitter: a path analysis," *Electron Markets*, vol. 21, no. 3, p. 161, Aug. 2011, doi: 10.1007/s12525-011-0065-z.

[6] K. Żbikowski and P. Antosiuk, "A machine learning, bias-free approach for predicting business success using Crunchbase data," *Information Processing & Management*, vol. 58, no. 4, p. 102555, Jul. 2021, doi: 10.1016/j.ipm.2021.102555.

[7] C. Lin and A. Kunnathur, "Strategic orientations, developmental culture, and big data capability," *Journal of Business Research*, vol. 105, pp. 49–60, 2019.

[8] D. Arora and P. Malik, "Analytics: Key to Go from Generating Big Data to Deriving Business Value," in *2015 IEEE First International Conference on Big Data Computing Service and Applications*, Mar. 2015, pp. 446–452, doi: 10.1109/BigDataService.2015.62.

[9] M. Delbaere, E. F. McQuarrie, and B. J. Phillips, "Personification in Advertising," *Journal of Advertising*, vol. 40, no. 1, pp. 121–130, Apr. 2011, doi: 10.2753/JOA0091-3367400108.

[10] R. J. Cohen, "Brand Personification: Introduction and Overview," *Psychology & Marketing*, vol. 31, no. 1, pp. 1–30, 2014, doi: https://doi.org/10.1002/mar.20671.

[11] A. Griva, C. Bardaki, K. Pramatari, and D. Papakiriakopoulos, "Retail business analytics: Customer visit segmentation using market basket data," *Expert Systems with Applications*, vol. 100, pp. 1–16, Jun. 2018, doi: 10.1016/j.eswa.2018.01.029.

[12] L. Molenaar, "Data-driven personas: Generating consumer insights with the use of clustering analysis from big data," *undefined*, 2017. /paper/Data-driven-personas%3A-Generating-consumer-insights-Molenaar/d9c8d7adb6d4c1c2ab1f7c95c202c6770879c57b (accessed Jun. 28, 2020).

[13] J. An, H. Kwak, S. Jung, J. Salminen, and B. J. Jansen, "Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data," *Social Network Analysis and Mining*, vol. 8, no. 1, p. 54, 2018, doi: 10.1007/s13278-018-0531-0.

[14] J. G. Carbonell, R. S. Michalski, and T. M. Mitchell, "An overview of machine learning," *Machine learning*, pp. 3–23, 1983.

[15] J. Alzubi, A. Nayyar, and A. Kumar, "Machine Learning from Theory to Algorithms: An Overview," *J. Phys.: Conf. Ser.*, vol. 1142, no. 1, p. 012012, Nov. 2018, doi: 10.1088/1742-6596/1142/1/012012.

[16] B. J. Jansen, S. Jung, and J. Salminen, "Creating Manageable Persona Sets from Large User Populations," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, United Kingdom, 2019, pp. 1–6, doi: 10.1145/3290607.3313006.

[17] J. D. Kelleher, B. M. Namee, and A. D'Arcy, *Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies*. MIT Press, 2020.

[18] A. E. Mohamed, "Comparative Study of Four Supervised Machine Learning Techniques for Classification," *International Journal of Applied Science and Technology*, vol. 7, no. 2, p. 14, 2017.

[19] M. E. Celebi and K. Aydin, Eds., *Unsupervised Learning Algorithms*, 1st ed. 2016 edition. Springer, 2016.

[20] D. D. Lee and S. H. Seung, "Learning the Parts of Objects by Non-negative Matrix Factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[21] R. T. Rust and M.-H. Huang, "The Service Revolution and the Transformation of Marketing Science," *Marketing Science*, vol. 33, no. 2, pp. 206–221, Jan. 2014, doi: 10.1287/mksc.2013.0836.

[22] T. Agrawal, *Hyperparameter Optimization in Machine Learning: Make Your Machine Learning and Deep Learning Models More Efficient*, 1st ed. edition. Apress, 2020.

[23] P. Probst, A.-L. Boulesteix, and B. Bischl, "Tunability: Importance of Hyperparameters of Machine Learning Algorithms," *Journal of Machine Learning Research*, p. 32.

[24] S. Thirumuruganathan, H. Rahman, S. Abbar, and G. Das, "Beyond itemsets: mining frequent featuresets over structured items," *Proc. VLDB Endow.*, vol. 8, no. 3, pp. 257–268, Nov. 2014, doi: 10.14778/2735508.2735515.

[25] C. R. Muh. Ibnu, J. Santoso, and K. Surendro, "Determining the Neural Network Topology: A Review," in *Proceedings of the 2019 8th International Conference on Software and Computer Applications*, New York, NY, USA, Feb. 2019, pp. 357–362, doi: 10.1145/3316615.3316697.

[26] R.-S. Wu and P.-H. Chou, "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach," *Electronic Commerce Research and Applications*, vol. 10, no. 3, pp. 331–341, May 2011, doi: 10.1016/j.elerap.2010.11.002.

[27] B. J. Jansen, K. Sobel, and G. Cook, "Classifying ecommerce information sharing behaviour by youths on social networking sites:," *Journal of Information Science*, Feb. 2011, doi: 10.1177/0165551510396975.

[28] B. Kitchens, D. Dobolyi, J. Li, and A. Abbasi, "Advanced Customer Analytics: Strategic Value Through Integration of Relationship-Oriented Big Data," *Journal of Management Information Systems*, vol. 35, no. 2, pp. 540–574, Apr. 2018, doi: 10.1080/07421222.2018.1451957.

[29] M. A. Hossain, S. Akter, and V. Yanamandram, "Revisiting customer analytics capability for data-driven retailing," *Journal of Retailing and Consumer Services*, vol. 56, p. 102187, Sep. 2020, doi: 10.1016/j.jretconser.2020.102187.

[30] T. H. A. Bijmolt *et al.*, "Analytics for Customer Engagement," *Journal of Service Research*, vol. 13, no. 3, pp. 341–356, Aug. 2010, doi: 10.1177/1094670510375603.

[31] A. Griva, C. Bardaki, K. Pramatari, and G. Doukidis, "Factors Affecting Customer Analytics: Evidence from Three Retail Cases," *Inf Syst Front*, Jan. 2021, doi: 10.1007/s10796-020-10098-1.

[32] A. Maté, J. Trujillo, and J. Mylopoulos, "Specification and derivation of key performance indicators for business analytics: A semantic approach," *Data & Knowledge Engineering*, vol. 108, pp. 30–49, Mar. 2017, doi: 10.1016/j.datak.2016.12.004.

[33] S. Thirumuruganathan, S. Jung, D. Ramirez Robillos, J. Salminen, and B. J. Jansen, "Forecasting the nearly unforecastable: why aren't airline bookings adhering to the prediction algorithm?," *Electron Commer Res*, Jan. 2021, doi: 10.1007/s10660-021-09457-0.

[34] J. Salminen, I. Kaate, A. M. S. Kamel, S. Jung, and B. J. Jansen, "How Does Personification Impact Ad Performance and Empathy? An Experiment with Online Advertising," *International Journal of Human–Computer Interaction*, vol. 0, no. 0, pp. 1–15, Aug. 2020, doi: 10.1080/10447318.2020.1809246.

[35] B. Denizci Guillet, "Online upselling: Moving beyond offline upselling in the hotel industry," *International Journal of Hospitality Management*, vol. 84, p. 102322, Jan. 2020, doi: 10.1016/j.ijhm.2019.102322.

[36] J. Jiang, M. Yang, M. Kiang, and A.-F. Cameron, "Exploring the freemium business model for online medical consultation services in China," *Information Processing & Management*, vol. 58, no. 3, p. 102515, May 2021, doi: 10.1016/j.ipm.2021.102515.

[37] P. W. Murray, B. Agard, and M. A. Barajas, "Market segmentation through data mining: A method to extract behaviors from a noisy data set," *Computers & Industrial Engineering*, vol. 109, pp. 233–252, Jul. 2017, doi: 10.1016/j.cie.2017.04.017.

[38] J. Wechsler and J. Schweitzer, "Creating Customer-Centric Organizations: The Value of Design Artefacts," *The Design Journal*, vol. 22, no. 4, pp. 505–527, Jul. 2019, doi: 10.1080/14606925.2019.1614811.

[39] Y. Watanabe *et al.*, "ID3P: Iterative Data-driven Development of Persona Based on Quantitative Evaluation and Revision," in *Proceedings of the 10th International Workshop on Cooperative and Human Aspects of Software Engineering*, Piscataway, NJ, USA, 2017, pp. 49–55, doi: 10.1109/CHASE.2017.9.

[40] L. Nielsen, *Personas - User Focused Design*, 2nd ed. 2019 edition. New York, NY, USA: Springer, 2019.

[41] B. Jansen, J. Salminen, S. Jung, and K. Guan, *Data-Driven Personas*, 1st ed., vol. 14. Morgan & Claypool Publishers, 2021.

[42] B. J. Jansen, J. Salminen, and S. Jung, "Data-Driven Personas for Enhanced User Understanding: Combining Empathy with Rationality for Better Insights to Analytics," *Data and Information Management*, vol. 4, no. 1, pp. 1–17, 2020, doi: https://doi.org/10.2478/dim-2020-0005.

[43] T. Mijač, M. Jadrić, and M. Ćukušić, "The potential and issues in data-driven development of web personas," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 1237–1242, doi: 10.23919/MIPRO.2018.8400224.

[44] J. Grudin and J. Pruitt, "Personas, Participatory Design and Product Development: An Infrastructure for Engagement," p. 8.

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

[46] J. Pruitt and J. Grudin, "Personas: Practice and Theory," in *Proceedings of the 2003 Conference on Designing for User Experiences*, San Francisco, California, USA, 2003, pp. 1–15, doi: 10.1145/997078.997089.

[47] L. Nielsen, "Engaging personas and narrative scenarios," PhD Thesis, Samfundslitteratur, Copenhagen, Denmark, 2004.

[48] L. Nielsen, K. S. Hansen, J. Stage, and J. Billestrup, "A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations," *International Journal of Sociotechnology and Knowledge Development*, vol. 7, no. 1, pp. 45–61, 2015, doi: 10.4018/ijskd.2015010104.

[49] Y. Chang, Y. Lim, and E. Stolterman, "Personas: From Theory to Practices," in *Proceedings of the 5th Nordic Conference on Human-computer Interaction: Building Bridges*, New York, NY, USA, 2008, pp. 439–442, doi: 10.1145/1463160.1463214.

[50] S. Faily and I. Flechais, "Persona Cases: A Technique for Grounding Personas," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2011, pp. 2267–2270, doi: 10.1145/1978942.1979274.

[51] F. Meissner and E. Blake, "Understanding culturally distant end-users through intermediary-derived personas," in *Proceedings of the South African Institute of Computer Scientists and Information Technologists Conference on Knowledge, Innovation and Leadership in a Diverse,*

Multidisciplinary Environment - SAICSIT '11, Cape Town, South Africa, 2011, p. 314, doi: 10.1145/2072221.2072266.

[52] J. Salminen, S. Jung, S. A. Chowdhury, S. Sengün, and B. J. Jansen, "Personas and Analytics: A Comparative User Study of Efficiency and Effectiveness for a User Identification Task," Honolulu, Hawaii, USA, Apr. 2020, doi: https://doi.org/10.1145/3313831.3376770.

[53] Revella, Adele, *Buyer Personas: How to Gain Insight into Your Customer's Expectations, Align Your Marketing Strategies, and Win More Business*. Wiley, 2015.

[54] Lauren Sorenson, "6 Core Benefits of Well-Defined Marketing Personas Lauren Sorenson," Dec. 13, 2011. https://blog.hubspot.com/blog/tabid/6307/bid/29583/6-core-benefits-of-well-defined-marketing-personas.aspx.

[55] D. Spiliotopoulos, D. Margaris, and C. Vassilakis, "Data-Assisted Persona Construction Using Social Media Data," *Big Data and Cognitive Computing*, vol. 4, no. 3, Art. no. 3, Sep. 2020, doi: 10.3390/bdcc4030021.

[56] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[57] A. Darliansyah, M. A. Naeem, F. Mirza, and R. Pears, "SENTIPEDE: A Smart System for Sentiment-based Personality Detection from Short Texts," *Journal of Universal Computer Science*, vol. 25, no. 10, pp. 1323–1352, 2019.

[58] H. Liu, L. Yao, Q. Zheng, M. Luo, H. Zhao, and Y. Lyu, "Dual-stream generative adversarial networks for distributionally robust zero-shot learning," *Information Sciences*, vol. 519, pp. 407–422, May 2020, doi: 10.1016/j.ins.2020.01.025.

[59] S. Jung, J. An, H. Kwak, M. Ahmad, L. Nielsen, and B. J. Jansen, "Persona Generation from Aggregated Social Media Data," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, Denver, Colorado, USA, 2017, pp. 1748–1755, Accessed: Sep. 24, 2017. [Online].

[60] J. An, H. Kwak, J. Salminen, S. Jung, and B. J. Jansen, "Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data," *ACM Transactions on the Web (TWEB)*, vol. 12, no. 4, p. 27, 2018, doi: 10.1145/3265986.

[61] B. J. Jansen, S. Jung, and J. Salminen, "From flat file to interface: Synthesis of personas and analytics for enhanced user understanding," *Proceedings of the Association for Information Science and Technology*, vol. 57, no. 1, Oct. 2020, doi: 10.1002/pra2.215.

[62] S. Jung, J. Salminen, and B. J. Jansen, "All About the Name: Assigning Demographically Appropriate Names to Data-Driven Entities," Virtual conference, 2021, [Online]. Available: http://hdl.handle.net/10125/71108.

[63] B. J. Jansen, S. Jung, and J. Salminen, "Capturing the change in topical interests of personas over time," *Proceedings of the Association for Information Science and Technology*, vol. 56, no. 1, pp. 127–136, 2019.

[64] E. K. Tang, P. N. Suganthan, and X. Yao, "An analysis of diversity measures," *Mach Learn*, vol. 65, no. 1, pp. 247–271, Oct. 2006, doi: 10.1007/s10994-006-9449-2.

[65] C. N. Chapman and R. P. Milham, "The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Oct. 2006, vol. 50, pp. 634–636, doi: 10.1177/154193120605000503.

[66] C. N. Chapman, E. Love, R. P. Milham, P. ElRif, and J. L. Alford, "Quantitative Evaluation of Personas as Information," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Sep. 2008, vol. 52, pp. 1107–1111, doi: 10.1177/154193120805201602.

[67] O. Nasraoui, J. Cerwinske, C. Rojas, and F. Gonzalez, "Performance of Recommendation Systems in Dynamic Streaming Environments," in *Proceedings of the 2007 SIAM International Conference on Data Mining*, 0 vols., Society for Industrial and Applied Mathematics, 2007, pp. 569–574.

[68] H. T. Hertzberg, G. S. Daniels, and E. Churchill, "Anthropometry of flying personnel-1950," ANTIOCH COLL YELLOW SPRINGS OH, 1954.

[69] S. Bernard, L. Heutte, and S. Adam, "Influence of Hyperparameters on Random Forest Accuracy," in *Multiple Classifier Systems*, Berlin, Heidelberg, 2009, pp. 171–180, doi: 10.1007/978-3-642-02326-2_18.

[70] D. A. Siegel, "The Mystique of Numbers: Belief in Quantitative Approaches to Segmentation and Persona Development," in *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, New York, NY, USA, 2010, pp. 4721–4732, doi: 10.1145/1753846.1754221.

[71] L. Laporte, K. Slegers, and D. De Grooff, "Using Correspondence Analysis to Monitor the Persona Segmentation Process," in *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, New York, NY, USA, 2012, pp. 265–274, doi: 10.1145/2399016.2399058.

[72] D. Kim, S. Hong, B.-J. Park, and I. Kim, "Understanding heterogeneous preferences of hotel choice attributes: Do customer segments matter?," *Journal of Hospitality and Tourism Management*, vol. 45, pp. 330–337, Dec. 2020, doi: 10.1016/j.jhtm.2020.08.014.

[73] M. Böttcher, M. Spott, D. Nauck, and R. Kruse, "Mining changing customer segments in dynamic markets," *Expert Systems with Applications*, vol. 36, no. 1, pp. 155–164, Jan. 2009, doi: 10.1016/j.eswa.2007.09.006.

[74] "The antecedents and consequences of customer-centric marketing | SpringerLink." https://link.springer.com/article/10.1177/0092070300281006 (accessed Apr. 20, 2021).

[75] W. J. Reinartz and V. Kumar, "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing," *Journal of Marketing*, vol. 64, no. 4, pp. 17–35, 2000.

[76] Ø. Helgesen, "Customer segments based on customer account profitability," *J Target Meas Anal Mark*, vol. 14, no. 3, pp. 225–237, Apr. 2006, doi: 10.1057/palgrave.jt.5740183.

[77] Z. Ren, Q. Shen, X. Diao, and H. Xu, "A sentiment-aware deep learning approach for personality detection from text," *Information Processing & Management*, vol. 58, no. 3, p. 102532, May 2021, doi: 10.1016/j.ipm.2021.102532.