Persona analytics: Analyzing the stability of online segments and content interests over time using non-negative matrix factorization

Bernard J. Jansen *,1, Soon-gyo Jung 2, Shammur A. Chowdhury 3, Joni Salminen 4
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

ARTICLE INFO
Keywords:
Data-driven personas
Big data
Social media analytics
Customer segmentation
Longitudinal analysis

ABSTRACT
Personified big data and rapidly developing data science techniques enable previously unforeseen methodological developments for longitudinal analysis of online audiences. Applying data-driven persona generation on online customer statistics from a real organizational social media channel, we demonstrate how personas can be deployed to understand online customer patterns over time. We conduct 32 monthly rounds of data collection of customer demographics and content consumption patterns on the YouTube channel of a major publishing organization posting thousands of items of content and then algorithmically generate 15 personas monthly. We analyze the data-driven persona for changes monthly, yearly, and lifetime (period). Results show an average 40% change in the personas, and 78% of the personas experience more change than consistency for topic interests. The implications are that organizations frequently publishing online content should employ automatic data collection and periodic persona creation to ensure their customer understanding is current. For this, algorithmic data-driven systems that leverage methods for persona creation are recommended.

1. Introduction
1.1. Evolving analytics landscape
Improving customer engagement is a persistent goal of organizations producing social media content (Saggi & Jain, 2018; Zerbonio, Aloini, Dulmin, & Mininno, 2018). To improve user engagement (Turgeman, Smart, & Guy, 2019), organizations are required to understand their customer’s preferences (Shao, Chandramouli, Subbalakshmi, & Boyadjiev, 2019); in other words, organizations must understand what content is most engaging. However, two issues make this goal difficult. First, online content preferences may vary largely among customer segments, and second, these preferences may change over time. Using existing tools and techniques to address these concerns is cumbersome, and mining customer insights from big data (Griva, Bardaki, Pramatari, & Papakiratikopoulos, 2018) remains one of the top challenges in the field (Pérez-González, Colebrook, Roda-García, & Rosa-Remedios, 2019; Vidgen, Shaw, & Grant, 2017). Understanding longitudinal customer behavior (Eliacik & Erdogan, 2018), in particular, is hindered by various challenges of presenting the temporal changes in online content consumption patterns (Bagher, Hassanpour, & Mashayekhi, 2017; Wu, Dong, & Tang, 2018) in an easily understandable manner. Even though various conceptual frameworks have been presented to investigate temporal customer data (Bijmolt et al., 2010), there is little empirical work demonstrating the applicability of these frameworks using real longitudinal user data and, especially, using data science algorithms and automated algorithmic data collection.

When modeling online customer behavior via analytics (Bertani, Bianchi, & Costa, 2020), an additional issue is the loss of empathy of representation when expressing customer data with graphs and tables as opposed to real human beings with real goals, motivations, and pain points, as suggested in the human–computer interaction (HCI) literature (Wright & McCarthy, 2008). Therefore, there are serious methodological hindrances for social media managers and other stakeholders to model online customers as data representations that are both quantitatively accurate and qualitatively empathetic (Akhtar, Ekbal, & Cambria,
machine learning, for inferring insights from quantitative online data automatically through periodical data collection from YouTube Analytics API and then uses a combination of non-negative matrix factorization and name-picture-quote database to retrieve online customer data, as well as serving as an ample demonstration of applying artificial intelligence (AI) techniques, specifically applied to a wide range of business processes and customer-facing decision-making tasks (Bijmolt et al., 2010), as well as serving as an ample demonstration of applying artificial intelligence (AI) techniques, specifically applied to machine learning, for inferring insights from quantitative online customer data.

For this research, we demonstrate the usage of data-driven personas – created by applying automated data collection and data science algorithms – for longitudinal analysis of online customers. Our research is positioned in the growing body of scholarly works on the benefits of automation of business processes and customer-facing decision-making tasks (Bijmolt et al., 2010), as well as serving as an ample demonstration of applying artificial intelligence (AI) techniques, specifically applied to machine learning, for inferring insights from quantitative online customer data.

The general concern we are addressing, the challenges in presenting big data in a user-friendly manner such that people in real organizations can infer valuable insights (France & Ghose, 2019) from it, has been widely noted by scholars in HCI (Chapman, Love, Milham, Elfri, & Alford, 2008; Matthews, Judge, & Whittaker, 2012), and marketing (Fernández-Gavidianes, Juncl-Martinez, García-Méndez, Costa-Montenegro, & González-Castaño, 2019; Germann, Lilien, & Rangaswamy, 2013; Xu, Frankwick, & Ramirez, 2016). Thus, our contribution provides value and applicability in many fields struggling with numerical online user data sense-making. We also contribute to the trending personification of big data (Stevenson & Mattson, 2019) by demonstrating a methodology with multiple benefits for researchers studying personas/online users and practitioners who desire human-like descriptions of their online customer segments customer-centric decision-making.

### 1.2. Personas over time

Personas are imaginary but realistic people that are created to assign a human face (Sproul, Subramani, Kiesler, Walker, & Waters, 1996) to a user, customer, or customer segment. This representation is typically instantiated in the form of a persona profile (Cooper, 2004) comprising the key attributes of the segment (Shin, 2017) represented by the persona (Nielsen, Hansen, Stage, & Billestrup, 2015). Personas have been integrated into commercial decision-making in many domains, including software and product development, marketing, and design (Dhawada, Greenstein, Gramopadhye, & Davis, 2007; Friess, 2012; Nielsen & Hansen, 2014), as personas are cognitively compelling for empathizing with users, customers, or customer members. Creators of personas have traditionally employed qualitative data collection approaches, including focus groups, interviews, or surveys. These collection approaches normally result in a handful of personas for a given organization. In contrast, data-driven persona creation draws from automatic programming interfaces and data science algorithms and has the potential to generate a larger set of personas on demand, thus representing the underlying user segments in a more thorough and time-sensitive manner (Salminen, Kwak, An, Jung, & Jansen, 2018).

Indeed, one of the central criticisms of personas is that creation using any method can stale over time, with the personas no longer representing the underlying population segment’s demographics or behavioral properties (Chapman & Milham, 2006). Seen as the underlying data of the personas being no longer valid (Chapman et al., 2008; Chapman & Milham, 2006; Salminen, Kwak, et al., 2018), the logical remedy is employing additional rounds of data collection and updating the personas accordingly. Given that the data collection techniques for persona creation have historically been manual, such an iterative data collection and analysis process is both time-consuming and expensive. Thus, decision-makers who use personas are left in uncertainty concerning whether or not the personas represent the current target segment. This limitation is particularly dire for organizations distributing content online because the customers of these organizations tend to be extensive and diverse, with the potential for evolving customer bases. Interestingly, despite this widely acknowledged criticism of personas in HCI and marketing, there is limited research that investigates whether personas do become outdated — and, if so, how often one should carry out additional data collection. Previous literature also lacks empirical tools on how to verify how personas change in the first place.

### 1.3. Research gap and key concepts

These are the concerns that we specifically investigate for customer personas through a data-driven approach, (a) examining the change in personas for a large news and media organization and (b) exploring the change in content interests of personas. Specifically, with customer personas, we were unable to locate any prior research showing that personas’ topical interests change, apart from a pilot study (Jung, Salminen, & Jansen, 2019). This research adds to the continuum of data-driven persona research that applies quantitative methods for creating personas that enhance decision makers’ understanding of their users, customers, or audiences (Jung, Salminen, An, Kwak, & Jansen, 2018; Salminen, Jansen, An, Kwak, & Jung, 2019; Zhang, Brown, & Shankar, 2016) (Table 1).

Adequate personas are personas that represent customer segments of an online audience population (see Fig. 1). We collect aggregated audience statistics for 32 months from a large international news and media organization’s YouTube channel, having hundreds of thousands of social media followers and containing thousands of pieces of online content. We generate 15 personas monthly from this data, following a fixed data-driven persona creation methodology (An, Kwak, Salminen, Jung, & Jansen, 2018), and we ascertain the personas’ three highest topical interests. We compare the personas and their topical interests by month, year, and the entire data collection period. The findings demonstrate there is persona change at all periods of comparison. Moreover, the topical interests of the personas also change, and topical interests can change rather rapidly.

We first review the related literature, then present research objectives, the methodology, and finally, the results. We end the manuscript with the implications and promising areas for future research in this area.
1.4. Research goals and objectives

The research goals are fourfold:

(a) Examine, using real audience data, whether audience change is captured and reflected in quantitatively created personas and, if they change, to what extent these changes occur.

(b) Scrutinize if and, if so, how much the topical interests change over time.

(c) Determine how the topical interest change is connected with persona set changes.

(d) Authenticate an algorithmic, data-driven methodology for examining changes in personas over time.

As such, we define four constructs addressing the concepts of both time and change, which are Persona Persistence, Persona Stability, Topic Persistence, and Topic Consistency and associated measures, as shown in Table 2.

From our research goals, we address three research questions (RQs):

- **RQ01**: What is the persistence of individual personas?
- **RQ02**: What is the (a) monthly, (b) yearly, and (c) period change in persona sets?
- **RQ03**: What is the topical consistency of individual personas?

The motivational foundation is to investigate changes over time in (a) a set of personas and (b) the topical interests of the personas. Given that customer segmentation is a well-known industry practice (Tynan & Drayton, 1987), we are also concerned whether or not there are changes in the topical interest of the customer segment, in addition to the
changes in the audience. This provides organizational insight into the use of personas for understanding customers (Lessmann, Haupt, Coussenent, & Bock, 2019). Thus, findings provide implications for organizations interested in monitoring their online audience personas for changes and forming strategic (e.g., C level) and tactical (e.g., implementation level) audience engagement plans accordingly.

2. Literature review

2.1. Change of personas over time

Although there are reported benefits for organizations employing personas (Goodwin & Cooper, 2009), including that they give a human face to impersonal analytics data, there are also concerns about their practical value. A serious and noted concern is that periodic data collections are needed to keep the personas updated (Mulder & Yaar, 2007) or to verify that the personas are still valid for practical decision-making purposes. This concern is notably present in the persona literature (Chapman & Milham, 2006) and is, by reason, derived from the premise that the user population, such as social media audiences (Zareie, Sheikhalamadi, & Jalili, 2019), experience periods of instability or change (Drutsa, Gusev, & Serdyukov, 2017) that should be reflected in the personas as they are provided to decision-makers.

Although a conceptually reasonable assertion (Mulder & Yaar, 2006), there is little empirical research investigating if personas do change over time or evidence that organizations need to update persona data to address changes (Drutsa et al., 2017; Viana & Robert, 2016; X. Zhang et al., 2016). Of the prior conceptual research that does exist, Chapman and Milham (2006) highlight that personas can go stale, but they offer no empirical support. Jais, Hignett, Allen, and Hogevorst (2016) discuss the need to update personas to represent different stages of dementia patients, but they do not carry out the updating, and the research did not focus on audience personas. Zhang et al. (2016) assert that data-driven personas can facilitate persona updating, but they offer no insights on when this updating is needed or how to do it. In a similar vein, using the support of data-driven personas, Mijac, Jadić, and Ćukušić (2018) note that personas cannot be easily updated, but they provide no empirical support for the statement that updating is even needed. Miasikiewicz and Luxmoore (2018) discuss a method for linking personas to actual user data for real-time updating, but this is again done at a conceptual level. Similarly, Nielsen (2019a), Nielsen (2019b) discusses the need to update personas but provides no information on whether or not personas change in real-world analytic datasets. Thus, the academic literature that addresses persona updating acknowledges this effort as conceptually important, but the same literature provides no empirical support that change does occur or tools for identifying if changes have occurred and how frequently.

The practitioner literature does go into slightly more detail. For example, Ritchie (2013) acknowledges the need to keep personas updated with fresh customer data. Eddine (2016) recommends updating personas at least yearly, offering no empirical evidence for this recommended update interval. Flaherty (2016) points out that one should update personas when there is a change in the user segment, although there is no guidance provided on what degree of change causes a change in personas or how to detect this change. However, Flaherty (2016) does report results from a survey of 156 professionals using personas. Twenty-eight percent of respondents reported having updated personas quarterly or more frequently, and these organizations also reported a higher impact of the personas than organizations that did not update. Adlin (2017) makes two counterclaims, both unsupported by empirical evidence, that personas either need updating rapidly or last nearly forever.

The concern for whether or not personas need to be updated is a key research gap to investigate, given that creating personas is not easy, cheap, or quick, as it may involve specific expertise and manual labor through ethnography sessions or focus groups (Goodwin & Cooper, 2009). Additionally, with these approaches, it may be practically impossible to compare personas from discrete datasets. Even when automated methods are employed (An et al., 2018; Jung, Salminen, Kwak, et al., 2018), effort and time remain in the creation process, as some steps cannot be completely automated (e.g., creation of training data for machine learning models, selection of the optimal number of personas and other hyper-parameters). Therefore, longitudinal analysis is important for detecting, for example, personas that are consistently appearing as audience segments (to understand loyal customers) or detecting personas that appear sporadically (to understand customers who come and go). Therefore, whether or not further data collection is needed has fundamental and applied implications, particularly for those organizations with big online populations, including content publishers with online customers whose preferences might rapidly vary (Sánchez & Bellogín, 2019).

2.2. Change in topics over time

Audience personas also have the concern of changing topical interests. The wide availability of content online permits individuals to designate their content preferences via likes, views, shares, and comments (Al-Aous, An, & Jansen, 2019; Lee & Tandoc, 2017). With increasing online content production, there are efforts in social media and other services (Wang, Yeoh, Richards, Wong, & Chang, 2019) to personalize online content to users/customers/audiences depending on their interests, requiring an emphasis on topical preferences from the audience groups (Li, Bai, Wenjun, & Xihao, 2019; Sánchez & Bellogín, 2019) or on personalization.

This is a critical issue for online content producers, including blog sites, social media services, media organizations, and news services. The use of accessible metrics and online data allows for the documentation of key content for the generation of personas (Jansen, Jung, Salminen, An, & Kwak, 2017). However, there is limited existing research that would show how changing customer interests can be captured and reflected using personas and keeping the interests alongside the personas as they evolve temporally. When leveraging personas in decision making, ensuring the audience personas are updated in terms of topical interests and changes in these topical interests is of critical importance for the generation of online content. For example, online content creators and social media managers can investigate the changing interests of a persona to guide their content production planning.

There has been a diversity of prior efforts in personalization aimed at the distribution of content online based on appreciating customer’s interests, which does change over time (Jiang, He, & Allan, 2014). For instance, Watters and Wang (2000) identify feature phrases of news content. Yet, it appears that online news audiences prefer a mixture of serendipitous content and personalization (Shepherd, Duffy, Watters, & Gugle, 2001), possibly demonstrating that audiences are either interested in novel types of content or fatigue when exposed to similar topics over an extended period. Consequently, research has focused on identifying audience inclinations temporally, focusing on similarities among audience members (Lv, Meng, & Zhang, 2017) or explicit content aspects, including entity identification (H. Zhang, Boons, & Batista-Navarro, 2019).

Personalizing the news delivery and other content online is contingent on the capability to predict audience preferences, with content tailored favored by users (Han, Yue, & He, 2015), although stated and genuine interests may differ (Sela, Lavie, Inbar, Oppenheim, & Meyer, 2015). Therefore, building personalized content systems necessitates detecting and tracking existing interests and also forecasting what online content users might be most interested in the future (Mele, Rainain, & Creslani, 2019) or at least what associated future content (Toraman & Can, 2017) a persona could prefer. While content-creating organizations endeavor to do this, research on news online has revealed that news content generally focuses on a Western perspective, with identifiable ongoing shifts towards Asian countries and Middle Eastern
news content (Segev, Sheafer, & ShenHAV, 2013).

Additionally, prior work (Salminen, Kwak, et al., 2018) has shown that media and public attention are alike but also dissimilar dependent on the granularity of the conducted analysis, with resilient regional similarities concerning media and related public attention and also a considerable quantity of countries where public attention and media attention differ by topical interest. In these instances, the country-specific outlets might ignore the people in these countries, and these audiences seek other available outlets to address their content desires and needs (Salminen, Kwak, et al., 2018). At the individual level, there is limited existing prior work that examines the diversity of and changes in the online consumption of content. Previous research has illustrated decreasing diversity in content consumption, with the consumption diversity correlated with the content available diversity (Zhang, Zheng, & Peng, 2017). Still, there is a noteworthy research gap in defining the change or drift in the audience’s online content topical interests. We propose that a fixed data-driven persona creation approach is applicable in online content consumption. This concern is a significant research gap to address, as additional rounds of data collection are needed to update or verify personas. The issue has real-world implications, especially for organizations with large user populations whose audiences might fluctuate rapidly and desire to use personas to introduce empathy from their analytics data. Audience segments might fluctuate both by their demographic attributes and by topical preferences. These concerns form the underlying motivations for this research in which we focus on audience personas for an organization over an extended period.

2.3. Summary of prior work

In summary, there is a dearth of empirical evidence, both in academic and practitioner prior work, either supporting or refuting the claim that personas require routine updating, partially linked to users’ changing interests, which is especially applicable in online content consumption. This concern is a significant research gap to address, as additional rounds of data collection are needed to update or verify personas. The issue has real-world implications, especially for organizations with large user populations whose audiences might fluctuate rapidly and desire to use personas to introduce empathy from their analytics data. Audience segments might fluctuate both by their demographic attributes and by topical preferences. These concerns form the underlying motivations for this research in which we focus on audience personas for an organization over an extended period.

A particular challenge for analyzing personas over time has been the manual persona creation approach. Because online audience segments may evolve and this change should be reflected in the personas, one needs a systematic way of creating personas to reflect such changes. To this end, we record and analyze the audience changes using online analytic data. Audience segments might fluctuate both by their demographic attributes and by topical preferences. These concerns form the underlying motivations for this research in which we focus on audience personas for an organization over an extended period.

To conduct this study, we employ data from an international producer of, among other media, YouTube content. Given the amplified availability of online audience data, there is the prospect to employ data-driven personas resulting from social media analytics data (Vecchion, Mele, Ndou, & Secundo, 2018). Personas, of all types, employing this data can be automatically created (Jung et al., 2017), and there is a system, Automatic Persona Generation (Jansen, Jung, & Salminen, 2019; Jansen et al., 2019), with this research using a different organization’s dataset, an extended data collection period, and integrated findings.

3. Methodology

3.1. Data collection

The research context of our study is a major international news and media organization that had more than 3.6 million subscribers and thousands of online content pieces during the data collection. As such, this organization exemplifies those entities that dispense content via major online social media services, such as news content producers, marketing and advertising firms, etc. These types of organizations characteristically have diverse and large audiences and produce varied content online, and their audience population may change over time. We gather a complete record of audience interactions by age group, gender, and country for each individual content using the YouTube Analytics API. We assemble audience data monthly for all channel content published, inclusive from November 2016 through June 2019. Note that the YouTube Analytics API offers measurements for each content piece and associated user data (e.g., country location, gender, age) at a combined level. The individual audience data is not available to maintain the privacy of the customers. This privacy preservation is an important benefit of the approach given the legislative environment (e.g., General Data Protection Regulation (EU GDPR)), where organizations are restricted from utilizing individual-level data. Our solution, thus, is the “individualization” of the widely available aggregated data.

3.2. Generating Data-Driven Personas

For repeatedly generating personas over the data collection period, we applied a data-driven persona generation approach called APG (see Fig. 2). As APG is extensively discussed in prior work (An et al., 2018; Jung, Salminen, Kwak, et al., 2018; Salminen, Şengün, et al., 2018), we only briefly present it here.

Relying on non-negative matrix factorization (NMF) (Lee & Seung, 1999), APG takes aggregated audience data and identifies unique viewing patterns (Sánchez & Bellogín, 2019), then associates these distinctive sets of viewing patterns with related demographic attributes. A distinctive behavior pattern refers to customers’ interaction with a set of products, with this association being a latent factor. NMF is a data dimensionality reduction technique previously used in fields such as recommender systems (Boratto, Fenu, & Marras, 2021; Zheng, Ding, Lin, & Chen, 2016). APG applies NMF to an interaction matrix constructed by APG from the view counts of the organization’s social media content, resulting in sets of distinctive content consumption patterns (see Fig. 3).

In Fig. 3, V is the g × c matrix of g customer segments (Gg, Gg, ..., Gg) and c contents (C1, C2, ..., Cc). When V is decomposed, W is a g × p matrix, H is a p × c matrix, and ε is an error term. In this case, p is the number of latent factors (behavioral patterns), which controls the customer behavior patterns resolution. As the number of latent factors increases, the customer behavior patterns are more fine-grained. The column in W is a basis for the segment, and the row in H is an encoding that consists of coefficients that combine with each basis and represent a linear combination of the bases.

With NMF, a column in H represents each of the common content consumption patterns. The coefficient, Hg, shows the importance of content, Cg, to explain the content consumption pattern, P(i.e., distinct customer interaction pattern). As mentioned, H shows a set of distinct content consumption patterns represented by a linear combination of customer interactions with content.

Leveraging the user data obtained automatically from YouTube Analytics using NMF, each distinctive behavior pattern is related to one to more than one set(s) of demographics by a weight assessment that encodes the strength of the relationship between the demographic group and the latent pattern. The demographic group with the maximum NMF weight is linked to a specific behavior pattern (Tai, Cambazoglu, Gullo, Mantrach, & Silvestri, 2017), as shown in Fig. 4.

These weights are, in a sense, an evaluation of the generation approach with the numerical values representing how strongly the demographic segment is associated with the latent behavior patterns. This data-driven approach has numerous advantages, one of which is...
that it is reactive to audience interactions with old and new content. This 
is significant as the data is cascading (i.e., current content gets fresh user 
interactions, and newer content is further added with no previous user 
interactions). The consequence of this process is the most distinct set of 
personas based on both behavioral attributes and demographic attributes 
of the underlying population. This results in minimal personas that 
characterize, yielding complete persona profiles (again, see Fig. 1).

3.3. Identifying topical interests of personas

During APG’s persona creation, each piece of content is also automatically 
classified topically, providing content interests for each persona. With the results from APG, we can recognize the specific content that a particular audience segment could be interested in before the content is published. In this section, we briefly explain the content features and how we employ them for persona creation. First, we delineate a matrix that holds the content features (i.e., video and views), and we then generate an alternative matrix that characterizes a link between content features and an audience segment. At this juncture, we employ a supervised learning approach for topic classification (Cunha et al., 2021), and we infer news topics from the set of content pieces (specifically, video titles and descriptions) retrieved from the organization’s social media channel for topical categories. We then calculate the uppermost topical interests for each of the personas by computing the Z-score for F\(_t\) for any current persona, and \(\sigma\) is the standard deviation of \(F\).

\[
F_t' = \frac{F_t - \text{avg}(F_t)}{\sigma}
\]

3.4. Calculating persona stability

To calculate the change in persona sets, we equate the set of personas first month-over-month (MoM) (i.e., the intersection with the prior month), then year-over-year (YoY) (i.e., overlap for one year prior), and finally, the overlap from the initial to the final data collection during the period (CoLP) (i.e., overlap between the end and the beginning of the period). The overlap coefficient metrics for persona persistence are defined in Equations (2)–(4).

\[
M_{0}M_{Y} = \frac{|P_{PM} \cap P_{CM}|}{|P_{PM}|}
\]

(2)

\[
Y_{o}Y_{ F} = \frac{|P_{FY} \cap P_{CY}|}{|P_{FY}|}
\]

(3)

\[
C_{FLP} = \frac{|P_{L} \cap P_{I}|}{|P_{L}|}
\]

(4)

A larger Z-score indicates that a persona has a greater likelihood to view the content for a given topic than another persona in the set. The consequence is a topic ranking for each persona (see Fig. 3). The bars to the left designate topics of most attentiveness, and the bars to the rightmost indicate topics of smallest interest by the persona, compared to the entire set of personas produced for that data collection period.

We carry out the same process to algorithmically generate the set of personas and the topical interests of the personas for each data period. Therefore, the results generated from an identical process allow for comparing both personas and topical interests for all months. We employ the same methodological style for each of the monthly datasets, producing fifteen personas for each iteration period. Fifteen personas are more than the usual number of personas (Pruitt & Grudin, 2003); we consider the greater number of personas sensible for those organizations with large online audiences. The monthly data collection and systematic analysis outcome are monthly sets of audience personas during the period. When we have a comprehensive sequence of 32 data sets, we then display the top fifteen personas for each monthly period represented by the audience segment.

We use the APG system to systematize this process and show the listings of personas (see Fig. 6, with Fig. 7 showing one persona provided for readability of a topical interests set). Note that the end-user of APG can click the personas listed in Fig. 6 to reveal the complete persona profile (see Fig. 1) at a specific point in time. This flexibility affords the end users to “zoom in and out” from the big picture of their audience into the individual audience segments that the personas represent.
trivial to compute the variation (i.e., the change for each period, MoMPC, YoYPC, CFLPC, is one minus the applicable metric, 1 – x, x ∈ {MoMPC, YoYPC, CFLPC}).

3.5. Calculating topical consistency

For topical consistency, we equate each persona’s top three topical interests month-over-month (MoMT) (i.e., intersection with the previous month). Then, we designated the uppermost three topics given that this is what is contained within the persona profiles; it is the most probable set to be used for audience analysis by the organization, with order not considered. Explicitly, the topic consistency metric is defined in Eq. (5):

$$\text{MoMT} = \frac{|\text{TPM} \cap \text{TCM}|}{|\text{TPM}|}$$  \hspace{1cm} (5)

where TPM is the set of topics in the previous month and TCM is the set of topics in the present month. In our particular case, |TPM| = |TCM| = 3, as the top 3 topics are always considered. After we have the intersection for this metric, it is trivial to compute the MoMT change, i.e., the MoMT change (=1 – MoMT) consistency value.

4. Results

4.1. Exploratory findings

During the 32-month data collection period, we generate 15 personas each month, resulting in 37 unique personas during the entire period. Demographic information for these 37 personas can be seen in Table 3, with the personas being predominately male (89.2%) and in the age group of 25–34 (70.3%). The personas are from a diverse group of countries, with 18.9% of them being from the USA.

We now turn to our research questions, explicitly addressing research questions 1 through 3, inclusive, which focus on persona stability.

4.2. Persistence of personas

Addressing RQ01 (What is the persistence of individual personas?), we computed the persistency of personas (i.e., of the 37 personas, how many times each appeared in the 32 data collection periods). Results in Table 4 indicate that the average persistency for personas was 13.0 appearances (SD = 12.5). The minimum appearance was 1, and the maximum was 32. The median was 7. On average, the persistency was 40.5%.

The distribution of changes is presented in Table 5. There were five stable personas (13.5%) during the 32 months (i.e., these five personas always occurred in the set of 15 personas) sets, with another two personas (5.4%) appearing in 30 and 1 persona (2.7%) in 29 of the 32 months. So, there is a base set of eight personas that signify the core audience for this organization over the total period.

4.2.1. Month-over-month persona change

Moving to RQ02a (What is the monthly persona change?), we calculated the MoMPC change. Results (see Table 6) indicate that the average MoMPC change was 2.6 personas-per-month (a standard deviation of 1.7). The minimum MoMPC change was 0, and the maximum MoMPC change was 8. The median change was 2. On average, the MoMPC change was 17.0%.

Concerning the distribution of MoMPC, the results (see Table 7) show that there was a single month with no MoMPC in the personas. Most of
the MoM$_{PC}$ were one, two, or three personas change from the fifteen personas, which is in line with the average.

4.2.2. Persona year-over-year change

Moving to RQ02b (What is the yearly persona change?), we calculated the YoY$_{PC}$. Results (see Table 8) indicate there was a YoY$_{PC}$ of 3 personas (20.0%) during Year 1 and then a YoY$_{PC}$ of 1 persona (6.7%) for Year 2. So, the YoY$_{PC}$ is in line (average of 2.0 personas) compared to the MoM$_{PC}$ (average of 2.6 personas), signifying that a portion of these personas is falling out of the 15 for a given month and then returning at some later point. To investigate if these averages held, we calculate 20 rolling yearly periods, finding that the average number of changed personas was 3.35 (22.3%) (Max = 8, Min = 1, Median = 3), indicating that the organization has gone through yearly periods of both changes and stability. We investigate this issue in greater detail upon examining the persona change over the given period.

4.2.3. Change in persona for the complete period

Addressing RQ02c (What is the change in personas over the entire period?), we computed the C$_{FLPC}$, with findings (see Table 9) showing that there were 7 personas that changed from the start to the conclusion of the collection period, with the number about cumulative for the yearly change in personas.

4.3. Change in topical interests

Returning to topical consistency, we first present overall findings.

4.3.1. Exploratory results on topics

Table 10 presents the summative topic distribution through the 32 months of collection. Based on their occurrence, the topics can be grouped into three broad clusters, which are a highly prevalent cluster (Research, Religion, Politics, Arts & Culture, Iraq), a mid-popular cluster (General News, War & Conflict, Business & Economy, Human Rights), and a less-popular cluster (Poverty & Development, etc.).

4.3.2. Topical consistency

Returning to RQ03 (What is the topical consistency of individual personas?), the topical interests of personas change, and these interests may change quite rapidly (see Table 11), with the average MoM$_{TC}$ being 52.3% (SD = 24.7%).

From Table 12, we can observe that the topical interests of the persona sets vary by data collection period, with every data collection period showing at least some change in topics relative to the prior month.

From Table 13, we can observe that every persona experienced some topical change during the data collection period.

We then examine the MoM$_{TC}$ for each of the 37 personas. As shown in Table 14, 78.4% of the personas experienced greater topic change compared to 18.9% that experienced greater consistency.

4.3.3. Change in personas and topics

For greater insights into these findings, we investigate the relationship between changes in the persona set and overall changes in topical interests. We study the personas ranked by stability and the corresponding topical consistency via a Spearman correlation determining the direction and strength of the monotonic relationship of the personas’ stability and topical consistency. A Spearman correlation indicates that
there is no significant correlation between topical consistency and persona stability. So, the persona topical interests are not affected by the persona set stability. The higher stability of a persona does not correlate with a higher consistency of the persona’s topical interests over an extended period.

4.3.4. Organizational content verification

An aspect that might have influenced the topical analysis is the range of content produced by the given organization, in that the content could have altered in a given period relative to what was produced previously. To examine this likelihood, we equated the aggregated consistency of topics MoMT for the content with the persona set. For comparison using a standard metric, we also calculate the Jaccard coefficient that measures two sets’ intersection and accounts for the randomness. In this case, the sets are Set 1 = the set of topics in the preceding month, and Set 2 = the set of topics in the present month. We use the first month as the standard. Eq. (6) displays the equation used for the Jaccard coefficient.

\[
J = \frac{N_c}{N_a + N_b - N_c}
\]

where
Table 5
Persona stability during the data collection period. No. of Appearances is the number of months that a given persona appeared in the data set (i.e., a measure of stability). The bold results represent the core audience of the organization's channel.

<table>
<thead>
<tr>
<th>No. of Appearances</th>
<th>Personas</th>
<th>% Persistence</th>
<th>% of Personas</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>5</td>
<td>100.0%</td>
<td>13.5%</td>
</tr>
<tr>
<td>30</td>
<td>2</td>
<td>93.8%</td>
<td>5.4%</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>90.6%</td>
<td>2.7%</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>84.4%</td>
<td>2.7%</td>
</tr>
<tr>
<td>26</td>
<td>2</td>
<td>81.3%</td>
<td>5.4%</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>78.1%</td>
<td>2.7%</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>59.4%</td>
<td>2.7%</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>53.1%</td>
<td>2.7%</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>46.9%</td>
<td>2.7%</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>37.5%</td>
<td>2.7%</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>31.3%</td>
<td>2.7%</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>25.0%</td>
<td>2.7%</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>21.9%</td>
<td>8.1%</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>12.5%</td>
<td>2.7%</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>9.4%</td>
<td>5.4%</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>6.3%</td>
<td>5.4%</td>
</tr>
<tr>
<td>1</td>
<td>11</td>
<td>3.1%</td>
<td>29.7%</td>
</tr>
<tr>
<td></td>
<td>37</td>
<td></td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Table 6
MoMPC showing the absolute values and percentages for average, standard deviation, maximum, minimum, and median.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Monthly Change</th>
<th>% of Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>2.6</td>
<td>17.0%</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1.7</td>
<td>11.4%</td>
</tr>
<tr>
<td>Max</td>
<td>8</td>
<td>53.3%</td>
</tr>
<tr>
<td>Min</td>
<td>0</td>
<td>0.00%</td>
</tr>
<tr>
<td>Median</td>
<td>2</td>
<td>13.3%</td>
</tr>
</tbody>
</table>

Table 7
Aggregated MoMPC showing the number of personas changes over the 32 months, along with the percentage of change from the set of 15 personas in a month.

<table>
<thead>
<tr>
<th>Number of Persona Changes</th>
<th>Occurrence</th>
<th>MoMPC</th>
<th>% of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1</td>
<td>53.3%</td>
<td>3.2%</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>46.7%</td>
<td>3.2%</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>33.3%</td>
<td>6.5%</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>20.0%</td>
<td>25.8%</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>13.3%</td>
<td>38.7%</td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>6.7%</td>
<td>19.4%</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0.0%</td>
<td>3.2%</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td></td>
<td>100.0%</td>
</tr>
</tbody>
</table>

Table 8
YoYPC change showing the absolute values and the percentages.

<table>
<thead>
<tr>
<th>June 17</th>
<th>June 18 YoYPC (%)</th>
<th>June 19 YoYPC (%)</th>
<th>Average YoYPC (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 (20%)</td>
<td>1 (6.7%)</td>
<td>2 (13.4%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 9
CFLPC showing absolute value and percentage.

<table>
<thead>
<tr>
<th>Beginning (Nov-16)</th>
<th>End (Jun 19)</th>
<th>CFLPC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>

Table 10
Topical distribution during the data collection period of 32 months for the monthly 15 personas.

<table>
<thead>
<tr>
<th>Topic</th>
<th>Occurrence</th>
<th>% of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research</td>
<td>197</td>
<td>15.1%</td>
</tr>
<tr>
<td>Religion</td>
<td>128</td>
<td>9.8%</td>
</tr>
<tr>
<td>Politics</td>
<td>111</td>
<td>8.5%</td>
</tr>
<tr>
<td>Arts &amp; Culture</td>
<td>106</td>
<td>8.1%</td>
</tr>
<tr>
<td>Iraq</td>
<td>101</td>
<td>7.7%</td>
</tr>
<tr>
<td>News</td>
<td>92</td>
<td>7.0%</td>
</tr>
<tr>
<td>War &amp; Conflict</td>
<td>85</td>
<td>6.5%</td>
</tr>
<tr>
<td>Business &amp; Economy</td>
<td>84</td>
<td>6.4%</td>
</tr>
<tr>
<td>Human Rights</td>
<td>80</td>
<td>6.1%</td>
</tr>
<tr>
<td>Poverty &amp; Development</td>
<td>37</td>
<td>2.8%</td>
</tr>
<tr>
<td>Environment</td>
<td>33</td>
<td>2.5%</td>
</tr>
<tr>
<td>Elections</td>
<td>31</td>
<td>2.4%</td>
</tr>
<tr>
<td>Korean Relations</td>
<td>30</td>
<td>2.3%</td>
</tr>
<tr>
<td>UK</td>
<td>24</td>
<td>1.8%</td>
</tr>
<tr>
<td>Social Media</td>
<td>23</td>
<td>1.8%</td>
</tr>
<tr>
<td>Health and Wellness</td>
<td>19</td>
<td>1.5%</td>
</tr>
<tr>
<td>US</td>
<td>19</td>
<td>1.5%</td>
</tr>
<tr>
<td>Refugees</td>
<td>18</td>
<td>1.4%</td>
</tr>
<tr>
<td>Police</td>
<td>17</td>
<td>1.3%</td>
</tr>
<tr>
<td>Sport</td>
<td>17</td>
<td>1.3%</td>
</tr>
<tr>
<td>Syria</td>
<td>16</td>
<td>1.2%</td>
</tr>
<tr>
<td>GCC</td>
<td>15</td>
<td>1.1%</td>
</tr>
<tr>
<td>Russian Affairs</td>
<td>15</td>
<td>1.1%</td>
</tr>
<tr>
<td>Corruption and Investigations</td>
<td>8</td>
<td>0.6%</td>
</tr>
</tbody>
</table>

Table 11
MoMTC change showing absolute values and percentages for average, standard deviation, maximum, minimum, and median.

<table>
<thead>
<tr>
<th>Metric</th>
<th>MoMTC</th>
<th>% of Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>52.3%</td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>24.7%</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>15</td>
<td>100.00%</td>
</tr>
<tr>
<td>Min</td>
<td>2</td>
<td>4.76%</td>
</tr>
<tr>
<td>Median</td>
<td>7</td>
<td>14.29%</td>
</tr>
</tbody>
</table>

Table 12
The number of occurrences of monthly topical changes, with MoMTC and the percentage of occurrence from the 32-month period. Note: Month 1 is the base month.

<table>
<thead>
<tr>
<th>Number of Topic Changes</th>
<th>Occurrence</th>
<th>MoMTC</th>
<th>% of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>3</td>
<td>100.0%</td>
<td>9.7%</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>93.3%</td>
<td>3.2%</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>86.7%</td>
<td>3.2%</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>80.0%</td>
<td>3.2%</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>73.3%</td>
<td>3.2%</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>66.7%</td>
<td>6.5%</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>53.3%</td>
<td>19.4%</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>46.7%</td>
<td>9.7%</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>40.0%</td>
<td>6.5%</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>33.3%</td>
<td>16.1%</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>26.7%</td>
<td>16.1%</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>13.3%</td>
<td>3.2%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>6.7%</td>
<td>13.4%</td>
</tr>
</tbody>
</table>

Nₐ = the number of topics in the preceding month
Nᵦᵢ = the number of topics in the present month
Nᵦᵢᵢ = the number of topics in the juncture of the previous month and the present month.

In Table 15, the typical topical consistency was 89.5% (10.5% change), and the average Jaccard coefficient was 85.62, signifying a large similarity among the topics published MoM₁. These coefficients also show that the changes in interests shown in the personas topical drift analysis were the consequence of changes in audience population,
Changes in the audience inclinations, or changes due to both, and they were not due to the accessibility of or variations in the organizational content.

5. Repeated experiments

In order to verify the replicability of the methodology with other datasets (Fortuna, Soler-Company, & Wanner, 2021), we applied the same approach to a different organization’s YouTube channel (Jansen et al., 2019; Jung et al., 2019), collecting audience data for the entire set of published videos monthly from October 2016 to September 2018, inclusive. The organizational channel had greater than 580,000 subscribers and thousands of pieces of online content during the study period. We applied the same algorithmic techniques for generating personas and identifying changes in topics.

As reported in Jung et al. (2019), the average MoMT change was 2.8 personas (SD = 1.8), and the minimum MoMT change was 0; the maximum MoMT change was 7, and the median change was 2. The MoMT change was 18.3% on average. Thus, there was a YoY change of four personas (26.7%) during the first year and a YoY change of three personas (20.0%) for the second year. Seven personas changed from the start to the end of the study period, with the numbers in line with the YoY changes. Concerning topical changes, as reported in (Jansen et al., 2019), the average MoMT change was 29.2% among the set of 15 monthly personas (SD = 27.2%). Also, the minimum MoMT change was 0.0%, the maximum MoMT change was 82.6%, and the median was 25.0%. The average consistency was 97.0% (3.1% change). The Jaccard coefficient average was 0.93, which indicates a high level of likeness among the topics published. The specific values differed between the two organizations, as expected, given the differences in the quantity of content. However, the methodological approach is generalizable to organizations of different sizes, as long as the demographic and behavioral metrics can be tracked.

6. Discussion and implications

6.1. Empirical analysis of personas over time

We demonstrate that personas can change over time, as evidenced by what the organization examined here, specifically audience personas generated from real-world data. We also present a practical approach to determining changes in persona sets for an organization and a systemic methodology for creating these data-driven personas, conceptually shown in Fig. 8.

The core implication is that organizations employing personas ought to regularly participate in data collection to identify possible changes in the audience population. While this data collection might be done via manual methods, it is more practical for comparative purposes to use automated approaches, as done in this research. The use of online real-time data and algorithmic practices for the creation of persona facilitates detection and assessment of changes in audience personas over time. We further demonstrated that the topical interests of audience personas also change. Concerning online content consumption, these research findings support and develop those reported in (Zhang et al., 2017), which showed a decrease in the consumption diversity of content (i.e., the people consumed content that tended toward uniformity over time). These two findings confirm a central notion in the persona literature that personas change, implying that they need to be updated according to the pace of change in the customer population data.

6.2. Scientific scrutiny of personas

This research also speaks to the commonly cited denunciation by Chapman and Milham (2006) that personas are outside the scope of systematic validation because their creation cannot be exactly reproduced. By enumerating and standardizing the persona creation process...
with algorithmically created personas, we demonstrate that personas can be conveyed into the realm of scientific examination. As an example of this, in this research, we demonstrate a method to quantify and measure persona changes (stability and consistency), presenting results that authenticate the need for updating personas or any other customer/user aggregations periodically that are based on changing analytics data. The innovative aspect of this research, the notion of analyzing the change of audience demographics and behaviors using the concept of personas over time, is an inventive approach only made essentially promising by the accessibility of large-scale analytics data that can be used for creating personas (AlSabban, Alorij, Alshamrani, & Alharbi, 2020; Spiliotopoulos, Margaris, & Vassilakis, 2020).

6.3. Practical implications for identifying changes in audience preferences

Along with the changes in personas discussed in prior work (Chapman et al., 2008; Chapman & Milham, 2006), we also investigate a subtle distinction concerning a change in content likings. We show that these topical changes are due to both (a) an underlying audience change and (b) portions of the existing audience experiencing changing topical preferences. For organizations producing online content, these findings highlight the need to monitor both demographic changes in the underlying audience and the shifts in preferences of audiences that impact content consumption monthly, as a rule of thumb. Practical use cases for APG in the context of online audience engagement include (1) identifying “loyal” personas that are consistently appearing as audience segments (What content are they drawn to? What demographics do they represent?); (2) identifying sporadic personas (How many segments appear from time to time? Is there an approach only made essentially promising by the accessibility of large-scale analytics data that can be used for creating personas (AlSabban, Alorij, Alshamrani, & Alharbi, 2020; Spiliotopoulos, Margaris, & Vassilakis, 2020).

6.4. Open avenues for future research

Future research may involve using the comparison approach to calculate the optimal number of personas in a period that could be the point where the change in the audience between data collection points is minimized. This would also address another open gap in personas research, namely, the “right number” of personas. Moreover, this analysis could be improved by investigating what the specific changes to the personas were and why these changes are significant for design decisions. Concerning content, research in the future could examine the pace of topical interest change for a high-frequency producer and a low-frequency producer of content. The supposition is that the personas (i.e., representing the composition of the audience) persist for low-frequency producers, as no new audience segments are attracted given that the changing content is stable. However, this supposition needs further investigation.

Other future research avenues involve determining the optimum number of topical interests in a given period. This could be the point where the change in the period is minimized, or the reaction by the audience segments to the content presented (Thielsch & Hirschfeld, 2019) is minimal. This also highlights a probable limitation in this research; we might overestimate topic change by restricting our focus to just three topics. Relatedly, our research only looked at audience personas and not segments from other domains (Li et al., 2019; Sánchez & Bellogín, 2019) on other platforms. Future research could address these limitations by examining other datasets and datasets from other platforms. We would expect the general trends to hold, although the specifics may change.

Moreover, one could also scrutinize the effect of distinguishing automatically topical shifts for content in order to integrate into a data-driven persona system. This research might assist content companies in circumventing content becoming non-novel or having audience members become uninterested with the content that is too similar to content consumed previously. Online publishers could anticipate new content that would excite the audience. An interesting developmental feature in APG is the prediction of how the personas are likely to engage with a given story based on the personas’ historical engagement with similar content. This feature, based on topic modeling and similarity scoring, is illustrated in Fig. 9.

It would also be interesting to evaluate the personas’ generation by
obtaining the latent factors’ values for a set of specific users and appraise whether these users’ demographic information was inferred correctly by the system. However, coordinating this naturistic study has several challenges. Finally, there is a need for a user study examining how decision-makers (e.g., social media managers, content producers, editors) perceive the data-driven personas and how the use of APG serves to improve their professional decision making. Despite there being strong logical reasoning as to why the system can provide value for real organizations, it is important that this is validated in the field.

7. Conclusion

It is confirmed in this research that audience personas and the topics they are interested in can shift over time, empirically confirming a criticism of personas that their apprising and validation requires the recurrent collection of new data. For the organizations that we examine here, there were extensive changes in personas within a comparatively short period. This research also confirms that the topical interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References


Coeper, A. (2004). The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity (2nd Ed.). Pearson Higher Education.


