

Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data

J. AN, Qatar Computing Research Institute, Hamad bin Khalifa University

H. KWAK, Qatar Computing Research Institute, Hamad bin Khalifa University

S. JUNG, Qatar Computing Research Institute, Hamad bin Khalifa University

J. SALMINEN, Qatar Computing Research Institute, Hamad bin Khalifa University

M. ADMAD, Strategy Division, Al Jazeera Media Network

B. JANSEN, Qatar Computing Research Institute, Hamad bin Khalifa University

We develop a methodology to automate the creation of imaginary people, referred to as personas, by processing complex behavioral and demographic data of social media audiences. From a popular social media account containing more than 30 million interactions by viewers from 198 countries engaging with more than 4,200 online videos produced by a global media corporation, we demonstrate that our methodology has several novel accomplishments, including (a) identifying distinct user behavioral segments based on the user content consumption patterns, (b) identifying impactful demographics groupings, and (c) creating rich persona descriptions by automatically adding pertinent attributes, such as names, photos, and personal characteristics. We validate our approach by implementing the methodology into an actual working system, and we then evaluate it via quantitative methods by examining the accuracy of predicting content preference of personas, the stability of the personas over time, and the generalizability of the method via applying to two other datasets. Research findings show the approach can develop rich personas representing the behavior and demographics of real audiences using privacy preserving aggregated online social media data from major online platforms. Results have implications for media companies and other organizations distributing content via online platforms.¹

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

1 INTRODUCTION

Personas are representations of users or customer segments presented in the form of an imaginary person. Personas are used in system development [Cooper 2004; Pruitt and Adlin 2005], product design [Goodwin and Cooper 2009; Smith 1956] and marketing [Revella 2015; Stern 1994], among many other fields and industry verticals. Personas are integrated into design processes and workflows [Dharwada et al. 2007; Eriksson et al. 2013; Friess 2012; Judge et al. 2012; Nielsen and Hansen 2014], for both long and short-term projects [Judge, Matthews and Whittaker 2012], with a reported positive return on investment in the business sector [Drego and Dorsey 2010]. Personas also relate to

¹ This manuscript is an extended version of earlier workshop and short conference papers or builds off prior work described in earlier papers that are cited in this manuscript.

analytics efforts from a variety of domains for identifying, constructing, and assessing groups of people (i.e., users, customers, audience, or market segments) in order to optimize some performance metrics (e.g., speed of task, ease of use, effectiveness of effort, sales, revenue, or engagement). Whereas market segments are typically presented as numbers or textual descriptors, personas are presented as fictitious people. This representation provides certain benefits compared to number-based analytics systems. Experiments in the field of psychology have shown that individuals lack the capacity to handle numbers in decision-making situations [see e.g., Kahneman and Tversky 1972]. In contrast, personas are intuitively understood because they characterize other human beings, thus evoking an individual's innate ability of empathy and immersion [Krashen 1984]. By substituting numbers with more approachable data representations, several benefits, including intuitive data representations, alignment of user understandings in the organization, and immersion of designers and developers in the user circumstances, can be achieved [Nielsen 2004]. Therefore, personas provide an alternative approach to analytics, emphasizing empathy and human attributes.

Although personas have claimed benefits beyond what data analytics by itself can provide [Adlin and Pruitt 2010; Beyer and Holtzblatt 1998; Dharwada, Greenstein, Gramopadhye and Davis 2007; Drego and Dorsey 2010; Eriksson, Artman and Swartling 2013; Friess 2012; Goodwin and Cooper 2009; Gułjónsdóttir and Lindquist 2008; Judge, Matthews and Whittaker 2012; Massanari 2010; Miaskiewicz et al. 2009; Pruitt and Grudin 2003; Rönkkö 2005], there are questions concerning the overall value of personas [Blomquist and Arvola 2002; Chapman and Milham 2006; Portigal 2008; Rönkkö 2005; Rönkkö et al. 2004] and the challenge of developing them [Chapman and Milham 2006; Guo et al. 2011; Nielsen 2004; Rönkkö 2005]. Additional challenges include verifying the accuracy of a persona [Pruitt and Adlin 2005] and clearly defining a usable persona [Grudin and Pruitt 2002; Marsden and Haag 2016].

Moreover, creating personas is typically not viewed as a cheap, easy, or quick process [Drego and Dorsey 2010]. Their construction has historically involved ethnographic methods, such as focus groups – therefore, the costs of a persona project can be tens or even hundreds of thousands of U.S. dollars, depending on the scope of the project. In addition, persona creation projects typically take months to complete. Also, as manual data collection is a onetime event, the personas created can become quickly outdated, resulting in the need for another round of data collection and increasing the cost further. Moreover, without real-time data, decision makers working in fast-pace industries have no confirmation whether the personas are representative of their current target users, audience, or customers. In other words, personas created in a conventional way cannot be matched with market segmentation due to lack of the real-time updating, and thus, in practice, it is hard to make strategic decisions that factor in the impact, revenue, or potential market of each persona. These limitations are especially acute for organizations that distribute content via major online platforms, such as modern media companies. With a potential audience in the millions, the traditional ethnographic methods lack scalability and cost efficiency in the online environment.

Addressing these limitations is the main driver for our research, in which we develop, implement, and evaluate an approach for leveraging aggregated data of user interactions with online content and then enhance the results with descriptive attributes to generate complete and rich personas profiles. The resulting persona generation process can

produce personas either as a standalone method or in conjunction with more traditional persona creation approaches, all while maintaining customer privacy. Therefore, our focus in this research is on the creation of personas in a cost-effective way using state-of-the-art computing techniques and substantial quantities of behavioral user data from online social media platforms.

2 LITERATURE REVIEW

Surprisingly, given the availability of large scale web analytics data, automatic generation of personas remains an open research question. Based on our review of literature, the use of behavioral online user data at a large scale is still missing from the persona research, greatly hindering the use of personas for analytical purposes.

Although the classical assumption is that personas are developed from data representing real users [Pruitt and Adlin 2005], this is not always the case, as behavioral user data is time consuming and expensive to gather [Nielsen and Hansen 2014]. However, the use of behavioral user data is crucial to make personas believable [Adlin and Pruitt 2010; Chapman and Milham 2006; Pruitt and Adlin 2005] and for designers and other decision makers, to appropriately leverage personas in their work. There are increasing recommendations to generate personas by using quantitative methods, for a variety of reasons [Mulder and Yaar 2006]. For example, prior research shows there is a need to ensure and empirically assess user data before assuming that personas actually describe real people [Chapman et al. 2008; Faily and Flechais 2011]. Otherwise, it becomes tempting to create personas that incorporate their creators' biases rather than beneficially informing the decision makers about users [Friess 2012]. In practice, though, attempts at developing personas from behavioral user data, of any size, have proven difficult due to, among other reasons, the time-consuming data collection, prohibitive costs, and freshness of manually collected user data [McGinn and Kotamraju 2008; Portugal 2008; Revella 2015]. Therefore, it has been noted that this key piece of using large scale user data is still lacking in the persona research and practice, despite the agreement that personas need to be based on such data [Chapman, Love, Milham, EIRif and Alford 2008; Rodden et al. 2010].

There have been some efforts in this area. Indeed, one can classify personas into three categories [Matthews et al. 2012] based on the level of data usage: (a) personas founded solely on data, (b) personas founded on data but with fictitious elements, and (c) fictitious personas developed without data. When real data is used, personas are typically developed from fieldwork, such as user interviews, direct observations, etc. [Clarke 2015; Cooper 2004; Goodwin and Cooper 2009; Grudin and Pruitt 2002]. The analysis methods utilizing this data depend in part on the size of the user data collected, although the analysis methods have been usually qualitative. A major critique of personas is that they are not based on sizeable quantities of first-hand user data, and there is not enough data for the application of quantitative methods [McGinn and Kotamraju 2008]. Therefore, the creation of personas from a quantitative, data-driven approach based on behavioral user data in sizeable quantities has remained an open research question to date [Rodden, Hutchinson and Fu 2010], with few efforts reported in the literature [McGinn and Kotamraju 2008], despite the increasing availability of web and social analytics data.

In addition, the research that has been conducted in transforming actual online user data into personas is limited. For example, [Jansen et al. 2011] used the data from

nearly 35,000 users of a social media platform to cluster users based on how they shared commercial information. However, the researchers did not use these results to generate personas, instead stopping at the cluster level, assigning descriptive names to each cluster. Chen, Pang, Xue [2014] used online user data to create mobility profiles, although they specifically were interested only in location. In another work, [Zhang et al. 2016] analyzed user-level clickstream data and extracted common sequences of clicks, identifying 10 common workflows using hierarchical clustering. They then presented five user facets based on the probability of platform use, which they then enhanced with a name to give it the fictitious facets. There has also been exploration in generating personas from publicly available Facebook data, with limited results, due to the privacy restrictions on the user profiles [An et al. 2016; Jansen et al. 2016]. In other domains, Guo, Zhu, Chen, Liu, Wu, and Guan [2011] leveraged social media data, again at the individual level, to develop credit risk profiles. In the marketing and advertising area, there is increased work on using large pools of online consumer data to segment markets [Smith 1956; Stern 1994], but there is limited prior work in creation of persona and personas descriptions in this domain.

The only available works on automatic persona generation that we are aware of include [An, Cho, Kwak, Hassen and Jansen 2016; An et al. 2016; An et al. 2017; An et al. 2017], Kwak et al. [2017] and Jung et al. [2017]. An et al. [2016] presented the first steps toward such goal by applying k-means clustering. However, that approach used individual-level data that is expensive to collect and has concerns regarding privacy. In a similar vein, Kwak et al. [2017] used k-means for clustering and found the limitation that a single demographic group must fall into one persona. However, in reality, various personas can be found from one demographic group, as people in the same demographic group can, and often do, behave differently. These limitations observed in our previous work provide the research gap for the current work, with initial results report in [An, Kwak and Jansen 2017; An, Kwak and Jansen 2017; Jung, An, Kwak, Ahmad, Nielsen and Jansen 2017].

Therefore, the creation of personas from a quantitative, data-driven approach based on actual, first-hand user data remains an open research question, motivating the current research.

3 RESEARCH OBJECTIVE AND NOVELTY OF RESEARCH

Our research objective is to develop a methodology for retrieving aggregated, privacy preserving user data from major social media platforms in order to identify distinct and impactful audience segments and then generate personas with realistic descriptions and attributes that represent these key user segments. This research is novel in several respects.

First, it is one of the first, research efforts to use online social media data at scale for automatic persona generation. This concept is novel in itself.

Second, the data from such platforms is already aggregated, unlike the limited prior work in identifying audience segments [Jansen, Sobel and Cook 2011; Zhang, Brown and Shankar 2016], or creating personas for the internal application of a private company [Zhang, Brown and Shankar 2016] using individual level data. However, due to privacy and other concerns, audience data from the major online platforms is not individualized. Instead, it has been aggregated, typically along coarse attributes, such as gender, complicating the generation of personas. Therefore, we must develop techniques to

decompose this aggregated data for persona generation while still respecting data privacy. The current research provides a solution using this aggregated data that is easy to collect through modern social media platforms, and it inherently overcomes the privacy issues due to its aggregated form.

Third, our approach is flexible in the number of possible personas generated. Typically, persona creation focuses on a small number of personas, from three to six, due to constraints imposed by qualitative methods. While appropriate for a traditional, small-scale data, a limited number of potential personas is not optimal considering the major social media platforms that are used by millions or even billions of users worldwide. As social media data sets are typically large and diverse, in the tens of millions of data points, the number of meaningful personas can be considerably high. By using our approach, we can leverage quantitative methods in generation and validation of our personas, and therefore capture the behavioral and demographic diversity of the userbase.

Fourth, our approach automates the persona generation process, from data collection to generation of persona descriptions, validating the inherent value and practicality of the proposed solution. Beyond our limited previous work [An, Kwak and Jansen 2017; Jung, An, Kwak, Ahmad, Nielsen and Jansen 2017], we could locate no prior research concerning generation of fully developed personas using aggregated data for those who distribute their products via major online platforms. In [An, Kwak and Jansen 2017], first steps were taken toward applying non-negative matrix factorization as the core technology for automatic persona generation. The research presented here greatly expands on this prior work, especially in the areas of methodology, generalizability, evaluation, and discussion of impact.

Fifth, our research approach demonstrates the case of digital content creators. There has been no prior research that we could locate concerning personas for digital content creators who distribute their content via social media channels, where the user data for persona generation begins with content consumption behavior (e.g., view, click on, download, scroll, etc.). The personas shown to content creators should capture the taste of the audience so that the content creators can publish more appealing content for the target personas. Understanding the different roles of personas for content creators is a unique aspect of this research. However, we postulate that the methodology presented here is generalizable to any organization that distributes content and products via major online platforms. Therefore, the potential impact of this research is broad and applicable across different domains and industry verticals.

4 AUTOMATIC PERSONA CREATION

4.1 Overview of the method

This section provides a general explanation of automatic persona generation (APG) methodology. In Section 5, we demonstrate how it can be applied to real user data from a major social media platform. To achieve our research objective of automatically generating personas of the user segments of a given social media account, our methodology consists of the six steps shown in Fig.1. The method is generalizable to any interaction pattern; however, in this research we use video views.

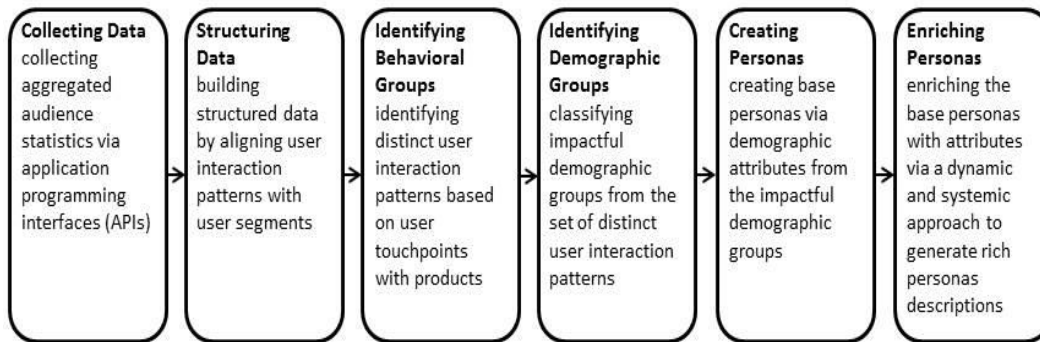


Fig. 1. Automatic Personas Generation (APG) Approach. The APG Approach is a Six-Step Process to Convert Raw Web Analytics Data into Rich Persona Descriptions.

In our approach to develop personas from aggregated audience statistics, first, we introduce non-negative matrix factorization to identify the number of significant behavioural patterns and impactful demographic groups that become the base of the personas. Next, we explain how to dynamically add personal attributes, which are user interests, demographics, photos, names, and other personal details, in order to build the rich persona descriptions that are shown to the end users of our system.

4.2 Data Collection and Structuring

4.2.1 General Alignment of Patterns of Interactions with Sets of Users

Once the data is collected from a social media platform, we first build a matrix representing users' interaction with the online products. We denote by \mathbf{V} the $g \times c$ matrix of g user groups (G_1, G_2, \dots, G_g) and c the online products (C_1, C_2, \dots, C_c). The element of the matrix \mathbf{V} , V_{ij} , is any statistic that represents the interaction of user group G_i for product C_j . For example, in the case of YouTube Analytics, V_{ij} is a view count for a particular video, C_j from user group G_i , which is defined by gender, age, and country, such as [Male, 25-34, South Korea]. In the case of data sources such as YouTube Analytics or Facebook Insights, V_{ij} is total minutes watched a particular video, C_j , from a user group G_i .

A user group (G_i) interacts with the set of digital products (C_1, C_2, \dots, C_c). So, a user group is defined as a set of the touch points with the collection of digital products. With this matrix (\mathbf{V}) as the basis, we can discover the number of significant latent patterns by decomposing it, which will become the basis of the personas, explaining the persona's preference toward particular products.

We note that a user group, G_i , can be an individual user if the data is available at that level of granularity and a privacy concern does not exist. This means that the research approach can be seen generalizable to user-level data, as well as the aggregated data used here. More specifically, this matrix approach is broadly applicable across 1) data of

different granularity and 2) any content type. In addition, beyond social media analytic tools, our approach can be applicable for any domain where the matrix V can be defined. For example, if an app store provides statistics concerning app downloads by different user groups, V_{ij} can be defined as the number of downloads from a particular user group for a particular app. In that case, our approach can find personas of that app store without any modification of the core algorithm presented here. Therefore, the algorithmic approach, as far as we can see, is generalizable.

4.2.2 Identification of Distinct Behavioral Patterns and Impactful Demographic Groups

Once we have the matrix V , the next step is to discover the underlying latent factors, or the product consumption patterns that become the basis of the personas. The matrix decomposition functions as shown in Figure 2.

$$\begin{array}{c}
 \boxed{\mathbf{V}} \\
 (g \times c)
 \end{array}
 =
 \begin{array}{c}
 \boxed{\mathbf{W}} \\
 (g \times p)
 \end{array}
 \begin{array}{c}
 \boxed{\mathbf{H}} \\
 (p \times c)
 \end{array}
 +
 \begin{array}{c}
 \boxed{\boldsymbol{\varepsilon}} \\
 (g \times c)
 \end{array}
 \quad (1)$$

Fig. 2. Overview of matrix decomposition for identifying distinct interaction patterns and then impactful demographic segments, which form the basis for the resulting personas.

From Fig. 2, V is our matrix of V the $g \times c$ matrix of g user groups (G_1, G_2, \dots, G_g) and c products (C_1, C_2, \dots, C_c). When decomposed, W is a $g \times p$ matrix; H is a $p \times c$ matrix, and ε is an error term. Here, p is the number of latent factors (behavioral patterns) that we can choose, which are used to identify unique sets of user interactions with products. The column in W is a basis for the personas, and the column in H is an encoding that consists of coefficients that combine with each basis and represent a linear combination of the bases. The resulting matrix decomposition equation is:

$$V = WH + \varepsilon \text{ or } V_{ij} = \sum_{k=1}^p W_{ik}H_{kj} \quad (1)$$

The basis and the encoding depend on what decomposition technique is employed. There are three widely used matrix decomposition methods for this purpose: principal component analysis (PCA), vector quantization (VQ), and non-negative matrix factorization (NMF) [Lee and Seung 1999]. In terms of actual technique, PCA, VQ, and NMF bring different decomposition results by having different constraints in W and H . With VQ, each column in H has to be a unary vector. In other words, only one entry in a given column in H has a non-zero value, and all others have to be zero [Gray 1984]. This constraint makes H too simplified to get meaningful behavioral patterns explained by a combination of content interactions. Therefore, VQ is not appropriate for our purpose.

With PCA, the rows in \mathbf{H} have to be orthogonal, and columns in \mathbf{W} have to be orthonormal [Jolliffe 2002]. PCA approximates an entry in \mathbf{V} as a linear combination of the corresponding row and the column in \mathbf{W} and \mathbf{H} , respectively. However, PCA entries in \mathbf{W} and \mathbf{H} can be either positive or a negative. These positive and negative coefficients lead to complex cancellations, making the results difficult to interpret. Therefore, PCA is not appropriate for our purpose.

By contrast, NMF does not allow negative entries in \mathbf{W} and \mathbf{H} . As no subtraction led by the negative coefficients is allowed, we consider a linear combination as only an additive combination of bases. This non-zero constraint makes interpretation of the matrix decomposition straightforward. In summary, we choose NMF to extract common content consumption patterns from the aggregated audience interaction statistics.

In NMF, a column in \mathbf{H} represents each of common content consumption patterns. The coefficient, H_{ij} , shows the importance of content, C_j , to explain the content consumption pattern, P_i . (i.e., distinct user interaction pattern). As mentioned, \mathbf{H} shows a set of distinct content consumption patterns represented by a linear combination of user interactions with content.

While there are other techniques besides matrix factorization that could be applied, such as clustering, which we attempted in prior work [An, Kwak and Jansen 2017], results from our previous work has shown NMF to be best suited for our purpose. In particular, clustering requires that there is only one encoding of behavioral pattern per a demographic group. This is a major drawback for personas based on aggregated user interaction data, as one demographic group can have multiple behavior patterns. given group (e.g., Women, 24-35, West Virginia) can include more than one behavioral pattern; thus, multiple persona can exist within the same demographic group. This also implies that the use of latent behavioral patterns has more discriminatory power than using demographic information because sub-segments of a demographic group may behave differently.

4.2.3 Creation of Base Personas

A row in \mathbf{W} represents each user group consisting of different common consumption patterns. The coefficient, W_{ij} , is a relative proportion of a consumption pattern, P_j , in a user group, G_i (i.e., impactful user demographic group). A row in \mathbf{W} represents how each user group can be characterized by different consumption patterns. A column in \mathbf{W} shows how a distinct consumption pattern is associated with different user groups. Thus, for each column, the user group with the largest coefficient can be interpreted as the most impactful user group for that corresponding pattern.

To determine the demographics of the representative user groups, depending on how the user groups are defined in \mathbf{V} , the most efficient way is to use the data broken-down by demographics when building \mathbf{V} . For instance, if \mathbf{V} has a row mapping into a group defined as [age group, gender, country], then it is trivial to find representative demographics of a persona. Social media analytic tools often provide user statistics in a format that we can leverage for a persona profile descriptive snippet.

We accomplish this in two-steps: (a) finding a representative user group for a persona, as outlined above and then (b) identifying the representative demographics of this group. Once we have identified the column in \mathbf{W} with the largest coefficient, we then select that demographic grouping from our data set. In our data collection, for example, YouTube

provides demographic classification of 2 genders x 7 age groupings x 198 countries (2,772 possible demographic grouping) per video, which is the ceiling of possible demographic personas that that can be addressed. Then, our approach enriches the base personas with other attributes.

4.2.4 Enriching the Base Personas

Following the methodology described above, we are able to elicit the age, gender and location for each behavioral segment, as well as their behavioral pattern, which becomes the basis of the personas. However, we need to add other information to build actual “imaginary people”; this information includes name, photo, and topics of interest, corresponding to a typical persona representation in the literature [Nielsen and Hansen 2014]. Since we desire the persona generation to be automatic, we employ a series of back-end databases and API accesses. To generate a name for a persona, we build a dictionary of names by collecting popular names by gender and year from each of the 181 countries. For example, there is information on the US and many other countries concerning the top 1,000 popular baby names for any year since 1879, often by ethnicity. Then, through [age group, gender, and country] of a representative demographic group, we can dynamically assign an age, gender, and ethnically appropriate name to a persona (see Figure 5(a)). For example, for a 25-year old female from the U.S., our system can assess one the common female baby names for females in the U.S. from twenty-five years ago for naming the persona.

Once we select the representative groups for each persona from **W**, we get (age, gender, country) information of the persona. Since the age group is represented as a range (e.g., 18-24) in YouTube, we choose a random value within it. Then, we randomly pick from the set of temporally appropriate names corresponding to the given demographic information from the database dictionaries we have built.

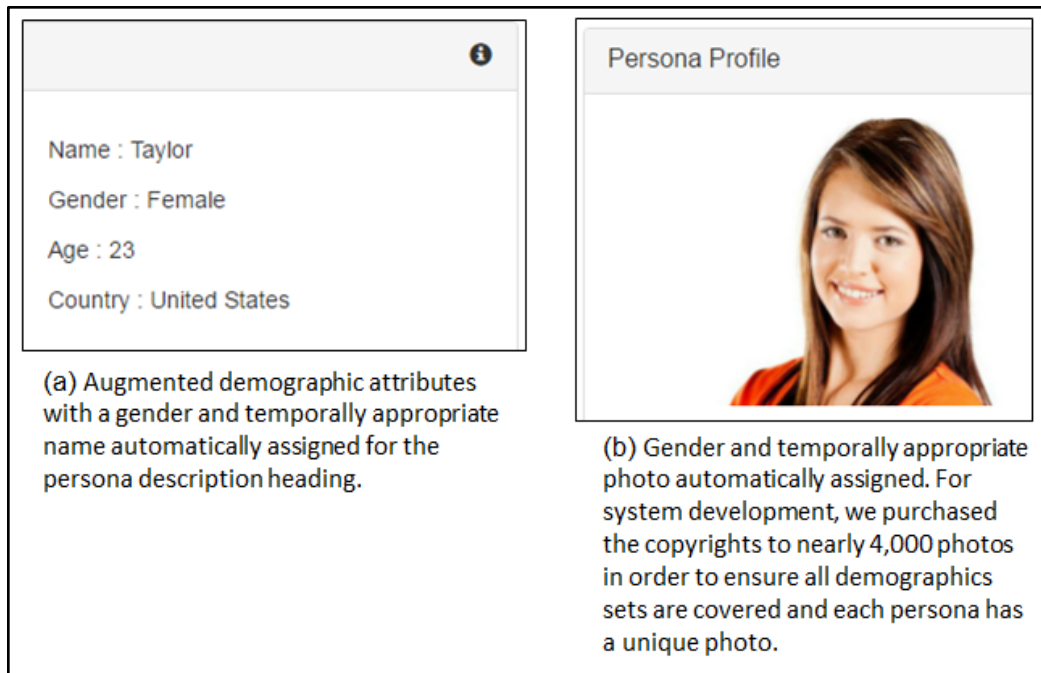


Fig. 3. Enriching of the persona descriptions with appropriate name and photo for the user segment.

To assign a photo to a persona, we purchase copyrights to nearly 4,000 commercial stock photos of models for different ethnicities, countries, genders, and ages. We ensure that we have photos for all possible personas, and we have multiple photos for some popular demographic groups, for which there may be multiple behavior patterns, which in our case, is the multiple video consumption patterns represented by a vector of videos viewed. Here, the selection of different styles of figures to represent different professions, interests, etc. can strengthen the expressive power of the persona, so we select varied photos for each and tag each photo with the appropriate metadata. Then, through [age group, gender, ethnicity, country, etc.] of a representative user segment, we assign an appropriate photo to a persona (see Figure 3(b)). So, if we need a photo of a twenty-three old white female from the U.S., we can automatically access that photo for our persona description using the corresponding metadata in the image database.


5 PRACTICAL DEMONSTRATION OF APPLYING THE APG METHODOLOGY

5.1 Overview

In the previous section, we described how the automatic persona generation works at a general level. In this section, we apply it to a real dataset originating from a social media account of a major news organization. The result, outlined in detail below, are rich insightful persona descriptions, generated automatically from aggregated, privacy

preserving social media data. Figure 5 shows one of the persona description outcomes of the described persona generation techniques in this case example.

Persona Profile



Name : Taylor
Gender : Female
Age : 23
Country : United States

About Persona

Taylor is a 23 year old female living in the United States and works in the Management field. She likes to read about US-affairs, Racism, and US-politics on her Mobile. She usually watches about 1.4 minutes of video.

Topics of Interest

More Interested Topics

- U.S.-affairs
- Racism
- U.S.-politics

Less Interested Topics

- Terror
- Environment
- Technology & Science

Quotes

"should be cheaper"

"Legalize it!"

"The sad truth in america its all about colors whites like whites and brown like browns period"

Comments

Add Comment

Submit

No Comments

Most Viewed Videos

- U.S.-politics** Is The South Racist? We Asked South Carolinians
- U.S.-politics** Donald Trump' In Mexico: We Asked Mexicans What They Thought About Trump
- U.S.-politics** Confederate Flag Debate - We Ask People In South Carolina What They Think
- News** Anonymous Declares War: On ISIS After Paris Attack
- Society** Legal Marijuana: How High Is Too High?
- Others** Cutting Hair With Swords And Fire
- U.S.-affairs** Riding The 'Death Train' Changed This Boy's Life Forever
- U.S.-affairs** The Life Of An Eight Ball Of Cocaine
- Others** Charlie Charlie Pencil Game
- U.S.-affairs** Disturbing Footage Of Mentally Ill Inmates Facing Abuse

Potential Reach

5,412,000 people

Gender (Female), age (18-24), country (United States), interests (Society, Anti-Racism, Politics), and language (English) based potential reach.

Fig. 5. Rich persona description, with traditional components and persona attributes, all generated automatically from online social media data. Additionally, as the system is online, one

can directly access the underlying data, providing credence to the personas [Matthews, Judge and Whittaker 2012].

In the next subsections, we describe the case company, data collection techniques, and persona generation.

5.2 Description of the case company

To develop and implement our approach, we leverage consumer user data from AJ+, an online news channel from the Al Jazeera Media Network. Two common goals for media organizations are to increase digital content consumption and enhance the user interaction with digital content. Specifically, with the news industry, prior studies point out considerable differences between production and consumption patterns of content [Abbar et al. 2015], as the online news industry is competitive and fluid [Abbar, An, Kwak, Messaoui and Borge-Holthoefer 2015; Kwak and An 2014]. Therefore, in the news area, as with many other verticals, proper understanding of customers is critically important, an issue which online audience data can address [Mao and Zhang 2015; Shuradze and Wagner 2016].

In this research, we collaborate with Al Jazeera (AJ+) for both data collection and analysis. We focus on the AJ+ YouTube channel² as the data source of the aggregated audience statistics, which we use as a proof of concept for our persona generation methodology research (see Figures 4(a) through 4(c)). We note that we do not lose generality in showing the proof-of-concept by using a single media account because we use the data provided by YouTube, which has a universal format for all YouTube channels. In other words, our system does not use any AJ+ dependent features; instead, we use the typical data that all YouTube accounts have. Also, the approach is transferable to other social media platforms, such as Facebook, that have identical or highly similar data variables.

² <https://www.youtube.com/channel/UCV3Nm3T-XAgVhKH9jT0ViRg>

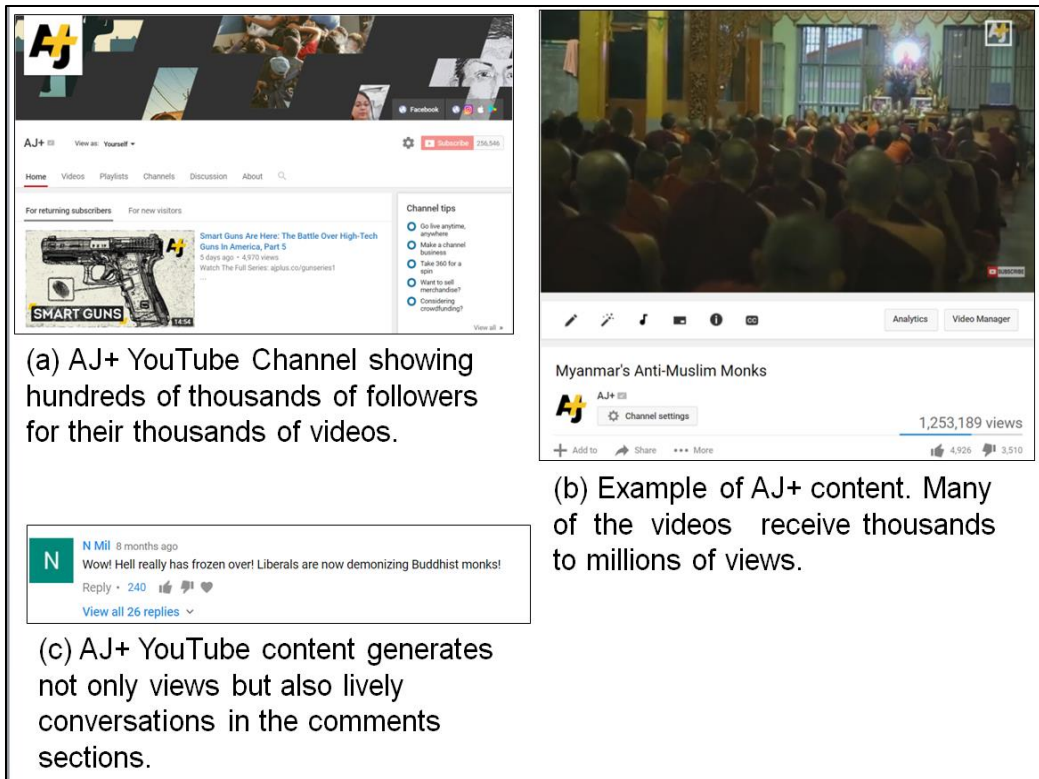


Fig. 4. AJ+ YouTube channel, sample digital content, and example from video comments section.

5.3 Data collection

For the owner of the channel, the YouTube API provides analytics data for each video and various user profile data, (e.g., gender, age, country location, and which site the user comes from), although at an aggregate level (see Fig. 5). Individual user data is not provided. Via the YouTube API, we collect the detailed record of product views by country, gender, and age group for each of AJ+ video. We focus in the research presented here on view counts due to their high volumes and being a key metric for digital content creators. A user group is defined by gender, age, and country, such as [Male, 25-34, South Korea]. We note that this detailed breakdown data is accessible only to an owner of a YouTube channel (i.e., AJ+). In summary, we collect data from 4,320 video products produced from June 13, 2014 to July 27, 2016 for the research presented here. Collectively, these videos have more than 30 million views from users of 181 countries at the time of the study.

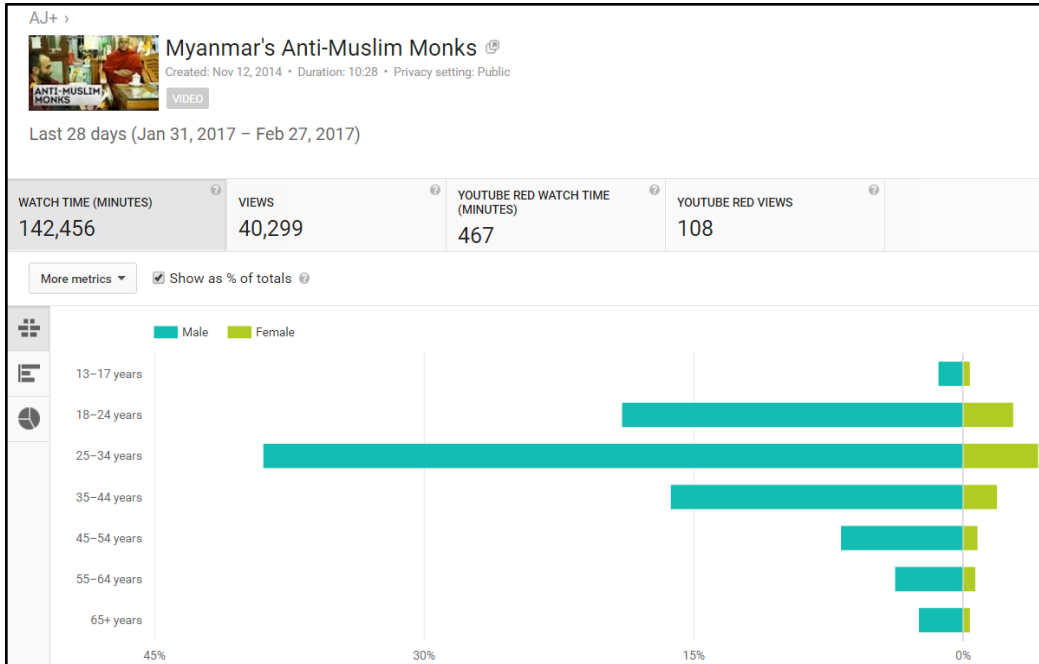


Fig. 5. Example of the AJ+ YouTube Analytics interface via the YouTube API, from which we access the research data.

5.4 Persona generation

The result of NMF is a set of base personas that encodes preferences towards each set of content products for a set demographic, analogous to consumer behaviors for a market segment. We now turn basic personas into rich personas by adding personal attributes to each, the last step in our approach, using the methodology as outlined in the following. The result of this is that we automatically generate personas based on actual user data from the AJ+ YouTube channel, a significant evolution of persona creation research.

We first identify behavioral patterns of AJ+ users by applying NMF and then adding personality to each of the resulting basic personas. To apply NMF, there are two implementation aspects to consider: 1) defining user groups and content in a matrix \mathbf{V} , and 2) choosing the number of personas to build. As discussed above, the YouTube channel analytics API provides the number of view counts of each video by demographic segments. We choose to use a combination of age, gender, and country as a group (G_1, \dots, G_g). A YouTube video naturally maps to a content (C_1, \dots, C_c). Then, the entry in \mathbf{V} , V_{ij} , is a view count of a content piece C_j by a user group G_i . In our dataset, we have 198 unique countries, two gender groups, and seven age groups, resulting in $198 \times 2 \times 7 = 2,772$ groups. We find that, among 2,772 groups, 2,491 (89.9%) groups' contribution to the view counts of 4,320 videos is extremely limited; across 4,320 videos, each of their total view counts is less than 1,000. In other words, each demographic group watches

each video 0.4 times on average. We exclude those groups in the following steps. We note that the resulting groups cover more than 95% of the total view counts across all the videos.

In choosing the number of personas to generate (or display via the system), we give users the flexibility to choose the number of personas generated. However, the maximum number of personas is constrained by the number of total groups ($|G|$), and it cannot be greater than $|G|$. When it is $|G|$, it will show the exact same user groups as YouTube provides. In the AJ+ case, users can choose any number of personas from 1 to the order of 10 (The condition for NMF is $|p| \ll \min(g, c)$). However, the cognitive load of tens of personas may make them unusable; therefore, in practice a smaller number is more reasonable, although the method and resulting system can generate as many personas as desired and as indicated by the data.

5.5 Persona Topics of Interest

We reveal the preferences of each persona from content consumption patterns, in particular, from a set of discriminative videos consumed in our AJ+ case study, although in practice it can be any product with which the user interacts. The discriminative videos of a persona are defined as the videos for which that persona has a higher probability to watch than another persona has. We identify the discriminative videos for each persona by Chi-square test [Casella and Berger 2001] on H at $p < 0.05$. From the Chi-square test, we can also rank the videos of a persona by calculating their effect sizes (ϕ) to find the most discriminative videos for each persona.

We examine the discriminative videos to summarize the user interest of personas. This summary provides insight into the usage goals of the users in engaging with AJ+ content. Topic analysis is widely used to abstract given content. To infer the topic of a video, we first attempted to use an automatic topic classification method Alchemy Taxonomy API [Alchemy Taxonomy API 2017; An, Cho, Kwak, Hassen and Jansen 2016], a popular text analysis service by IBM). However, from a follow-up manual evaluation, we found the method failed to identify the correct topics for the vast majority of the videos. For example, Alchemy classifies the video, “Fashion Models For A Right Cause,” simply as a “Fashion”, but the video is really about discrimination and human rights. The main reason for such failure is that the titles of AJ+ videos are very concise, more like catch phrases, often avoiding using direct descriptive topical words. The content descriptions are also concise and seemingly difficult to automatically classify by topic.

To address this problem, we manually classify a training set of videos by topic. Since there is no explicit taxonomy of topics for AJ+ videos, we first develop a topic classification scheme by conducting a qualitative content analysis. Qualitative content analysis is a series of flexible research methods to interpret the textual data through a systematic classification process [Cha et al. 2007]. Conventional content analysis begins with obtaining a sense of the whole by reading the data repeatedly and noting first impressions to derive codes, called open coding [Tesch 1990].

Table. 1. Topics and keywords used for classifying AJ+ videos.

Classification Topic	Sample of Keywords Set
Entertainment	NFL, sport, cooking, restaurant, cat, film, Messi

Environment	climate, crab, whale, tornado, wildfire, energy, recycle
Human-interest story	last with, motocross, emotional, tears, hug lady, cancer
International Affairs	Brexit, Scottish, Iran, UK, France, Kosovo, North Korea
Israel-Palestine	Palestine, Israel, Gaza
Racism	blacklivematter, fashion models, race, Ferguson, white
Refugees	refugee, Syria, Weiwei, overboard
Religion	hijab, muslim, rabbi, islam
Society	teacher, sexism, animal-right, activists, assault, HIV
South America	Mexico, Mexican, Venezuela, Brazil, Cuba
Technology and Science	Alpha Go, robot, wheelchair, Wikipedia, VR, lightning
Terror	bomb, explosives, ISIS, attack
US-politics	Obama, Trump, Clinton, Sanders, GOP, PACS, conventions
US-affairs	shooting, gun, NSA, Orlando, cop, KKK, native American

Table 1 lists the resulting 14 topics, along with a few example keywords describing the topics. Following an open-coding method, we identify the main topics of AJ+ videos in a two-phase process. We first read the titles and descriptions of 100 videos to develop an initial coding scheme and then used an affinity diagramming technique [Beyer and Holtzblatt 1998] to iteratively develop a classification scheme for topics until new topics did not emerge. The individual authors manually classified all videos into one of the topic categories. Once each researcher independently coded the videos, we then iteratively compared and re-coded the videos as necessary until we came to agreement of the classification schema. In addition, three researchers independently labelled a sample of 100 video headlines to test the robustness of the classification. The computed Fleiss Kappa was at 0.837, indicating an almost perfect agreement [McHugh 2012].

With the topic classification results, we get the fraction of videos of the topic t for a persona i (denoted as F_t^i). Then, using this fraction value, we examine which video topics a persona has a higher tendency to watch relative to the other personas. We quantify such topical preferences by computing the Z-score for F_t^i . For the topic t and the persona i , the Z-score can be computed as follows:

$$Z_t^i = \frac{F_t^i - \text{avg}(F_t)}{\sigma} \quad (2)$$

Where F_t is a set of F_t^i for any existing persona X , and σ is the standard deviation of the F_t . The higher Z-score means that a persona is more likely to watch videos of the topic than the other personas do.

Table 2. Topic preferences of each persona (T1: US-affairs, T2: US-politics, T3: Entertainment, T4: Racism, T5: Terror, T6: Environment, T7: Tech. & Sci., T8: South America, T9: Society, T10: Refugees, T11: Human-story, T12: Israel-Palestine, T13: Intl. Affairs, T14: Religion) represented by Z-value.

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14
P0	1.1	-0.5	0.2	1.0	-0.9	-1.3	-0.2	0.5	-0.3	0.3	0.6	0.2	0.1	-1.2
P1	0.4	1.5	-0.6	0.7	-1.0	0.3	-0.1	-0.7	0.5	-0.6	0.2	-0.2	-0.9	1.0
P2	-0.2	0.1	-0.1	1.2	-1.1	-0.6	-0.2	-0.4	-0.1	-1.2	2.0	1.5	1.3	0.5
P3	-0.5	-0.8	1.5	1.4	0.9	-0.3	-0.4	0.2	-1.2	-0.5	0.3	-0.1	-0.8	-0.4
P4	0.8	0.8	1.2	-0.3	-0.4	-0.3	-0.2	-0.1	-1.1	-0.3	-0.3	-0.3	0.8	-0.3
P5	-0.7	-0.1	-0.7	-1.1	-1.1	-0.8	-0.1	1.2	1.3	1.5	1	-1.5	-1.4	0
P6	1.2	0.3	0.6	-0.4	1.4	-0.8	-0.5	-0.8	-0.4	0.1	-0.1	-0.4	0.5	0.6
P7	-2.3	0.2	0.3	-1	-0.1	1	-0.6	-1	1.2	2	-1.2	0.9	0.7	-1.2
P8	0.5	-2.3	-2.2	-1.7	1.6	2.4	3	2.2	-1.3	0	-1.2	-1.5	-1.4	-1.2
P9	-0.3	0.9	-0.1	0.2	0.7	0.2	-0.6	-1	1.4	-1.2	-1.2	1.4	1.2	2.1

Table 2 shows the Z-scores of fourteen topics across the ten personas generated from the method described earlier. We notice that all ten personas have different distributions of Z-scores across the topics, indicating that they have unique topical preferences. As one example, Persona P3 has high positive Z-scores for Entertainment (T3), Racism (T4), Terror (T5), and South America (T8). In our content analysis, we observed that videos classified as Terror often mentioned the attacks by ISIS in Europe and videos classified as Racism focused on reporting the protests triggered by the racism incidents in the U.S. We summarize the user interests of this persona as following world news and seeking new information regarding recent terror attacks and protests against racism. Similarly, we summarize common interests (e.g., set of digital content topics preferred) of each of the other personas.

We filter users based on topical interests (see Figure 6(a)) by leveraging the collection of content viewed by that persona (see Figure 6(b)). Once identified, we can use this to associate to other online behaviors of the persona. For example, if a persona watches many videos about soccer, it is a reasonable assumption that users whose tweets are mainly about soccer in Twitter potentially have a similar content consumption pattern. However, in our approach we are interested in the positive cases that occur, and these matches do provide enough comments from similar users for the purpose of persona enrichment, given the previous assumption. We can leverage like demographics and

topic interests to get social media comments about the online products that we can incorporate into the persona description (see Figure 6(c)). By leveraging the underlying videos of these topical interests, and matching them with other user data, we can incorporate specific social media comments related to the topics (see Figure 6(d)). In the current version of the system, these social media comments are updated every time the persona description is displayed.

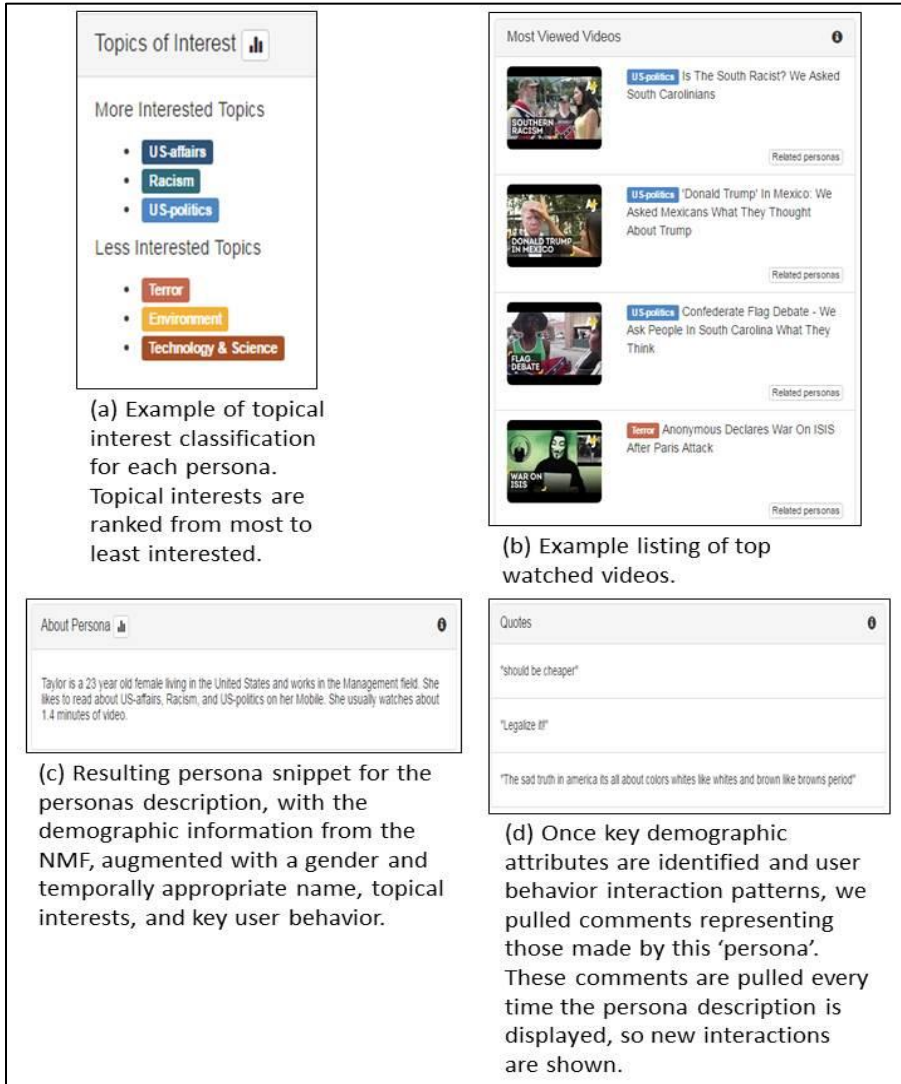


Fig. 6. Topics of Interest, Most Viewed Content, Persona Snippets, and Social Media Comments Used in the Persona Descriptions. Quotes are Social Media Comments from the Appropriate Platform (e.g., YouTube, Facebook, Twitter).

6 QUANTITATIVE EVALUATION OF AUTOMATIC PERSONA GENERATION METHODOLOGY

As one method to evaluate our approach for automatic persona generation, we predict the view counts of demographic groups underlying the personas, a combination of (age, gender, country), that form the base personas for new videos by using a ten-fold methodology on the data set.

6.1 Predicting Persona Interest for New Content

One of the benefits of using NMF for generating personas is a clear association, represented in $\mathbf{H}(p \times c)$, between personas and interests, or non-interest, in digital content. Beginning with this association, we can identify target personas of new content, H_n , even before content publication.

The problem of predicting interest in new content has been studied in recommendation systems. The most intuitive solution is to find similar content relative to the new content and approximate that the level of interest in similar content will remain the same. To compute the similarity of content in a robust way, we define content features. The features can be anything: topics, length, mood, color, price, and so on. Formally, we define a matrix, $\mathbf{F}(c_{\text{contents}} \times f_{\text{features}})$, capturing the features of content. We then can derive another matrix, $\mathbf{K}(p_{\text{personas}} \times f_{\text{features}})$, that represents an association between a persona and content features:

$$K = k(\mathbf{H}, \mathbf{F}) \quad (3)$$

Where k is a kernel function. Thus, we can rewrite equation (3) with some appropriate mapping function φ :

$$K = \varphi(\mathbf{H})\varphi(\mathbf{F}) \quad (4)$$

For computational simplicity, here we assume $\varphi=I$. In other words, the interest in content is the sum of the interest in its features. Then, we can get a direct multiplication of two matrices:

$$K = \mathbf{H}\mathbf{F} \quad (5)$$

By multiplying \mathbf{F}_{right}^{-1} for both sides, we get:

$$\mathbf{H} = \mathbf{K}\mathbf{F}_{right}^{-1} \quad (6)$$

where $\mathbf{F}\mathbf{F}_{right}^{-1} = \mathbf{I}$

The representation of \mathbf{H} in equation (6) guides us how to predict \mathbf{H}_n . For new content, we can define \mathbf{F}_n that represents new content and their features. By substitute \mathbf{F}_n into equation (6), we can get \mathbf{H}_n :

$$\mathbf{H} = \mathbf{K}(\mathbf{F}_n)_{right}^{-1} \quad (7)$$

$(\mathbf{F}_n)_{right}^{-1}$ can be computed by the following:

$$(\mathbf{F}_n)_{right}^{-1} = \mathbf{F}_n^T(\mathbf{F}_n\mathbf{F}_n^T)^{-1} \quad (8)$$

Equation (8) is valid when \mathbf{F}_n has linearly independent rows ($\mathbf{F}_n\mathbf{F}_n^T$ is invertible). If not, we split a set of new content products into several sets so that \mathbf{F}_n of each set has linearly independent rows. This procedure avoids the loss of generality of the method.

By combining equation (7) and (8), we write equation (9), representing the association between personas and new content:

$$\mathbf{H}_n = \mathbf{K}\mathbf{F}_n^T(\mathbf{F}_n\mathbf{F}_n^T)^{-1} \quad (9)$$

The key of equation (9) is that \mathbf{K} , the matrix representing an association between personas and features, does not need to be changed for newer content because \mathbf{K} depends on content features, not the content itself.

This is an application and advantage of our automatic persona generation methodology relative to other limited approaches that have been attempted for online data-driven audience profiling methods, especially for content creators. Not only providing personas of their audiences, our approach also helps content creators identify the target personas of new content once content features are selected and measured. The content creators then have an opportunity to refine their content prior to its publication to more directly appeal to the audience whom they want to target.

By combining equation (1) and (9), for new content, we get \mathbf{V}_n :

$$\mathbf{V}_n \cong \mathbf{W}\mathbf{H}_n = \mathbf{W}\mathbf{K}\mathbf{F}_n^T(\mathbf{F}_n\mathbf{F}_n^T)^{-1} \quad (10)$$

Similar to equation (9), it is possible to predict the views of new content by persona based on content feature of the new product.

6.2 Experimental Setup

Using this approach, we first define a training and a testing data set. We divide all 4,323 videos, which are ordered by published time, into 10 slices. Among the 10 slices, we use the 10th slice as our testing set, which is the latest 432 videos. Then, for training, we use some of the remaining slices to consider the recency and their expressive power.

Although there is a general belief that more training data leads to better prediction performance in machine learning, in our case, more videos to train the model is not necessarily helpful because the user base might change and evolve over time. In such cases, more data but data that is older might not reflect the behavioral patterns of the current users.

To understand better how the size or the recency of the training set affects the prediction performance, we iteratively run an experiment with varying number of slices in the training data from one (the most recent) to ninth (the oldest). For clarity sake, we use the percentage of the testing data to the whole instead of the number of slices; $N = 10\%$ means the ninth slice only, and $N = 30\%$ means the 7th, 8th, and 9th slices. We get 9 different sizes of the training data sets by changing N from 10% to 90%, with offset of 10%. For each training set, we construct a matrix \mathbf{V} , and by applying NMF, we get a matrix \mathbf{W} and \mathbf{H} . Then, we build a Latent Dirichlet Allocation (LDA) [Blei et al. 2003] topic model with 100 topics for each training set to construct a matrix \mathbf{F} and \mathbf{F}_n . In each training set, a video is represented as a vector by those 100 LDA topics. Table 3 shows five example topics that we find in one of our training set. Once we construct these five matrices, we estimate the view counts of new videos for demographic groups, \mathbf{V}_n (g groups \times n videos), according to equation (10).

Table. 3. The 5 topics (out of 100) identified by LDA algorithm

	Topics	Label
1	activist, palestine, israelis, shell, white, seattle, fair, house, together, cosby, syrian, fears, bill, robots, fears, france, many, celebrating, -sky, save	Israel-Palestine
2	refugee, crisis, rice, martin, muslim, teens, east, middle, domestic, ray, disease, largest, adrian, schooled, luthier, jr., king, asking, professor, radio	Refugees
3	debt, latino, elephant, america, reform, crisis, cables, celebrations, photojournalist, immigration, rome , student, trump, can't , graphic, spy, obama, israel, place, ted	Immigration
4	charleston, shooting, mother, victims, vigil, friday, joins, biker, barricades, protest, crimes, black, important, difference, war, kong, hong, u.s., artists, isil	US-Affairs
5	clinton, hillary, texas, court, finally, pakistan, commercials, shutdown, garland, u.s., forward, found, egyptian, weekend, sanders, shooting, protests, mass, don't	US-Politics

6.3 Measure of Evaluation

For each of the new videos, we rank demographic groups based on weight values in V_n . We compare this ranking with the true ranking of groups computed from the real view counts of that video by Kendall rank correlation coefficient. Since we have 432 test videos, we have 432 Kendall's coefficients (t) for each experimental run.

For comparison to a baseline, we employ two other models: 1) random model and 2) collaborative filtering (CF) model. The random model ranks groups randomly for a new video. The CF model computes the average view counts of each demographic group and uses them for any new video, as CF-based recommender system assigns an average behavior of users for the new content, which would represent what one would consider a standard web analytics approach. We, again, note that our goal here is not achieving the state-of-the-art prediction performance. Rather, our aim is to show that our persona well incorporates the topical video consumption patterns, which are identified by NMF, and such patterns are reliable enough to predict video consumption patterns for unseen videos. Thus, we use basic baseline models here, even more state-of-the-art models for future research.

6.4 Evaluation Results

Figure 7 shows the result of the experiment: the average t of cases where the result is statistically significant ($p < 0.05$). If two ranked lists are random, Kendall's coefficients can result in not significant. For fair comparison, in Figure 7, we only use those statistically significant cases among 432 test cases.

We first find that the average t for the random model is near 0.0 for any size of training set. The inferior performance of the random model proves that the view counts from each demographic group for the videos are far from the random construction. In Figure 7, our model outperforms CF based model in ranking the groups for new videos when N is 20% to 90%, and it shows comparable performance when N is 10%. These results demonstrate that our persona prediction performs very well at forecasting interest in new content by personas, and our approach outperforms, from 20% to 90%, the widely-used CF approach, even with a set of limited features.

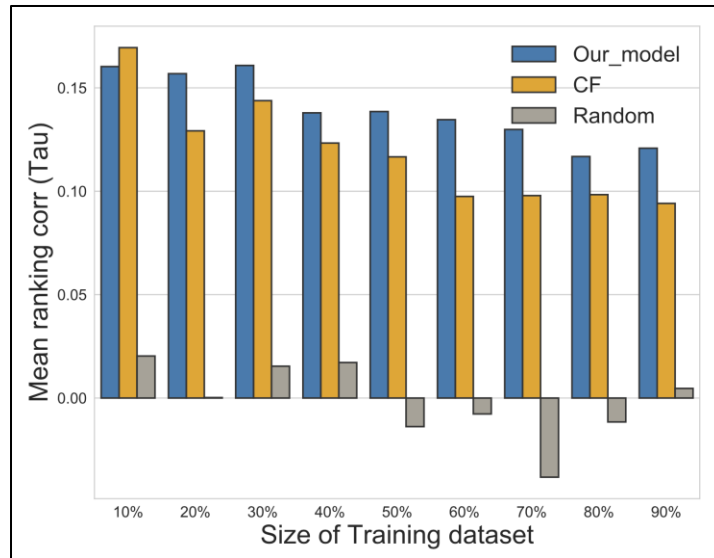


Fig. 7. The result of predicting the ranking of demographic groups by Random, CF, and our model. Y-axis value is the average of Kendall's rank correlation coefficient of our 432 test cases. X-axis value is the percentage of dataset used for training. 10% means we use the last 10% of dataset before the test dataset.

Even though our model overall outperforms Random and CF models, the improvement seems to be marginal. However, this is mainly because we do not consider those non-significant cases when computing the average of Kendall's coefficients. To show in how many cases Random and CF models result in "meaningful" ranked lists, Figure 8 depicts the number of the significant cases in each experiment for Random, CF, and our model.

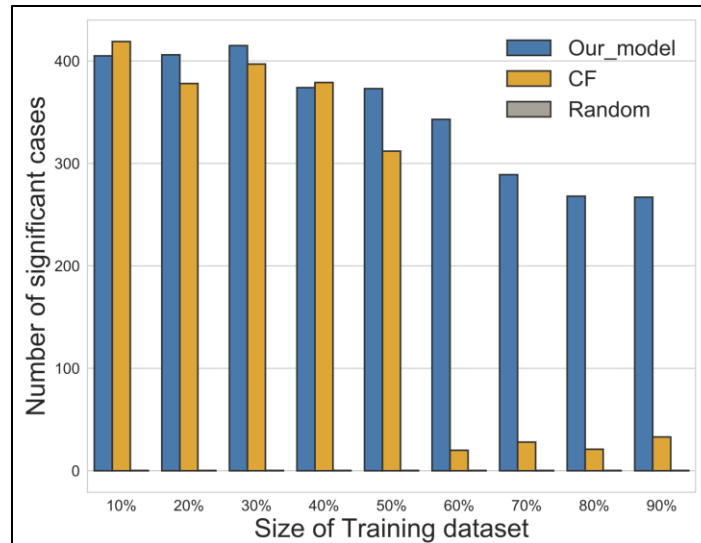


Fig. 8. The number of significant cases when testing Kendall's Rank Correlation Coefficients among 432 test cases by Random, CF, and our model. Y-axis value is the number of cases where the Kendall's rank correlation coefficients are significant among 432 test cases. X-axis value is the percentage of dataset used for training. 10% means we use the last 10% of dataset before the test dataset.

For the random model, there are not many significant cases for the random model. Also, for the CF-based model, the number of significant cases strikingly drops, from 412 significant cases to 18, when N is greater than 50% (Figure 8). From our stability analysis, we show that AJ+ encountered a sudden change of their consumer audience in that period. The CF-based model is not robust for such sudden changes, resulting in having no significant cases when $N > 50\%$. The average t of the CF model would show a significant drop when $N > 50\%$ if we plot the average t of all cases.

In contrast, our persona based approach is robust enough to adapt to the changes of the audience, as shown by the number of significant cases in Figure 8, even when N varies from 10% to 90%.

6.5 Stability of Personas

An aspect that can influence the generated personas is the recency of the user behavior data, which determines the size of the audience statistics data. In order to generate useful personas, we were interested in this recency effect on our persona generation methodology. Our personas represent the audience of online content produced, which might be dynamic and change over time. It is known that there are periods of stability and change in the audience of online platforms [Drutsa et al. 2017]. Thus, what is the appropriate time window is for the data set to both cover rich user behavior but also remain stable should be carefully considered.

To determine the appropriate time window for data slicing, we examine the stability of the personas content consumption patterns over time, as it is this content consumption

patterns that become a basis of the personas. We again sort all videos based on their published dates, then divide them into 10 subsets, each containing 10% of all videos. For each subset, we apply NMF, as defined above to identify ten basic personas within each subset. We then consecutively compare how similar or dissimilar are two sets of basic personas of consecutive subsets. In this respect, we are getting a temporal evolution overall of all personas during the data collection period.

Since each subset contains a different set of videos, the subsets are not necessarily generating the same set of basic personas. Thus, for comparison, we need to find a pair of matching personas. For that, we use the $K(p \times f)$ matrix in equation (4), which represents an association between a persona, p , and content features, f , commonly defined across all personas. For every pair of consecutive subsets, we compute the cosine similarity of any pair of 20 personas and find the 10 best matching pairs. Having high similarity values for all ten pairs indicates the audience behavior in two consecutive period subsets are similar. Low similarity values indicate the personas change over time.

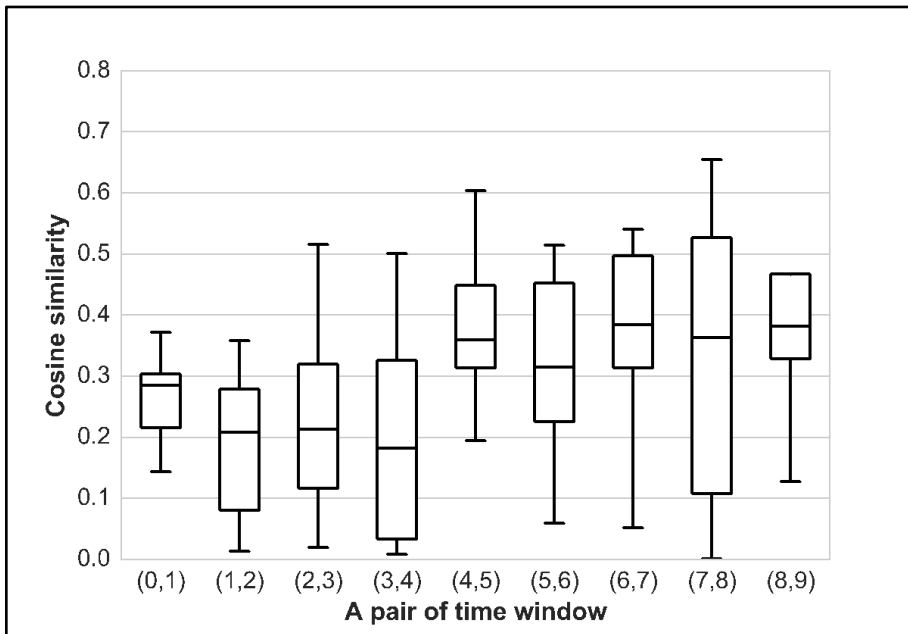


Fig. 9. Stability of basic personas over time. Each box plot presents the cosine similarity of the ten best-matching persona pairs between consecutive time windows.

Figure 9 shows the cosine similarity of the ten best-matching persona pairs as a box plot for every two consecutive subsets. During the first half of the period (reaching up to 50% of the videos), we observe the low cosine similarity, indicating that the audience behavior changes frequently. Then, there is a sudden spike in cosine similarity at 50% to 60% of videos, then keeping the same level of similarities, indicating the audiences are less changing in the second half of the period. The difference is also statistically significant as shown by an unpaired t-test ($t(3.27) = 78.58; p < .005$). The spike is

aligned to the period when AJ+ was experiencing a dramatic increase in popularity and intensely growing its audience. In that period, AJ+ attracted a large viewer base, reaching millions of views for their videos. We conjecture that the audience of AJ+ substantially changed at that point in time and so also did the personas. Since that time point, the persona set remains relative stable, indicating that our persona generation approach is both (a) valid for making product and content decisions aimed at the target user groupings, and (b) it is also reactive to changes in the consumer audience. Capturing this change of audience preferences or composition is not feasible with conventional persona generation methods without additional data collection, showing an advantage of our web analytics approach relative to qualitative methods.

6.6 Generalizability

To demonstrate the generalizability of our approach, we apply our method and create personas for two other media channels: AJ English and AJ+ Arabic. While these two are under AJ network, the target audience are different. Unlike AJ+, AJ English is a news service with a worldwide focus, and AJ+ Arabic focuses on 34 Arab speaking countries. Like we did for AJ+, we collect aggregate audience statistics from each of their YouTube channel and run the same analysis that we perform on AJ+ data. As YouTube Analytics provides the audience statistics in the same format, there is no need to modify the code for collecting, structuring, and decomposing the data. Results are shown in Table 4, illustrating that unique persons can be generated by our methodology from disparate data sets. In Table 4, note the divergence in personas, especially with age and location, among the three datasets.

Table. 4. Results from Analysis of Three Diverse Datasets. Results Illustrate the Generalizability of the Methodological Approach. Personas are Listed as *name, gender, age, country*. The Personas Capture the Behavior of Users in Different YouTube Channels, and They Were Easily Identified Using Our System.

Persona	AJ+ Personas	AJ English Personas	AJ+ Arabic Personas
P0	Robert, Male, 27, U.S.	Joshua, Male, 27, U.S.	Omar, Male, 29, Saudi Arabia
P1	Tyler, Male, 19, U.S.	Aarav, Male, 25, India	Nalkah, Female, 30, Jordan
P2	Benjamin, Male, 35, Portugal	Elijah, Male, 25, United Kingdom	Naila, Female, 32, Israel
P3	Ahmad, Male, 25, Malaysia	Ning, Male, 29, Philippines	Jahmir, Male, 28, Iraq
P4	Michael, Male, 49, U.S.	Alejandra, Female, 27, U.S.	Hala, Female, 33, Saudi Arabia
P5	Rade, Male, 26, Serbia	Jordan, Male, 33, Ethiopia	Youssef, Male, 34, Morocco
P6	Sarah, Female, 33, U.S.	Niran, Male, 62, Malaysia	Rashid, Male, 22, Saudi Arabia
P7	Adit Male 33 Indonesia	Goro, Male, 20, India	Mahmud, Male, 20, Jordan

P8	Manuel Male, 31, Mexico	Jamal, Male, 33, Zimbabwe	Shahin, Male, 40, Saudi Arabia
P9	Mei-ling, Female, 24, Taiwan	Elon, Male, 28, Nigeria	Pari, Female, 31, Turkey

Personas, by definition, are representations of the audience of a particular organization. Therefore, they do not “generalize” because each organization has a distinct audience. However, as shown, the methodology used to generate these personas does. As long as the data format follows the common structure provided by social media platforms, we are able to automatically generate the personas. By ‘data format’, we are referring to bucketed demographic data (e.g., [Women 24-35, West-Virginia]) associated with a performance metric (e.g., view count, click count, etc.). The user data is commonly structured this way by social media platforms. For example, the data format is always the same for each YouTube channel that exists. One can also see that the methodology generalizes beyond YouTube data to other social media data - for example, by changing the performance metric from ‘view count’ to another interaction metric (e.g. click, scroll, download, purchase), we can generate personas that capture different types of user behavior.

7 DISCUSSION AND IMPLICATIONS

In this research, we demonstrate several important results and implications concerning persona generation. First, the research shows that privacy-preserving aggregated online user data at scale from major social media platforms can be used to identify meaningful user segments and then automatically generate stable personas from these results. As such, this research addresses numerous shortcomings of current persona domain, such as lack of ties to actual user data, slowness of generation, and inability to easily update the personas, as reported widely in the previous literature [Chapman, Love, Milham, Elrif and Alford 2008; Chapman and Milham 2006; Faily and Flechais 2011; Marsden and Haag 2016]. Second, the research greatly advances persona and social media analytics research by presenting an approach to use online data at a magnitude for quantitative analysis and leverage this data to continually update the personas. By so doing, it also links two disparate streams of research, personas and social media analytics, addressing the issue of presenting analytics data in a meaningful way for easy communication and use by end users and organizations that heavily depend on this information. Third, our methodology is sensitive to changes in the user segments, both demographic and behavioral, which conventional persona creation methods are not.

Fourth, the methodology is conceptually and practically generalizable, as we demonstrate by running the analysis on additional datasets. Fifth, our research is novel in that it focuses on the use of data from major online social media platforms that are, in most cases, aggregated. The prior work of user profile generation focuses on individual user data [e.g., Guo, Shamdasani and Randall 2011; Zhang, Brown and Shankar 2016], to which many content creators do not have access. Our research using NMF demonstrates that one can use this aggregated data to both identify distinct and impactful user segments and then automatically generate rich personas while preserving the privacy of the individual users.

Sixth and finally, this research is also original in that it focuses on the increasingly common conundrum of digital content creators distributing their products to an extremely large, heterogeneous user base via major online platforms, which is becoming the *de facto* technology of distribution for digital products. While prior work has recommended a small number of personas often stemming from necessity, as manual persona generation does not scale to the datasets online companies are dealing with, this is not feasible or realistic for content, systems, or platforms with millions of worldwide users. In this research, we demonstrate an approach to generate a relatively large number of personas, while providing the end users with the choice of number of personas being shown through the user interface of the system. Thus, ours is among the first steps in focusing persona research on the needs of the users of analytics systems struggling with adopting human-based data representations while dealing with large and diverse datasets.

Although the current work represents a major step forward in automatic persona generation, there are many areas for future research. Given the reliance on streams of social media data, we could certainly implement direct access to the user data for the content creators, as suggested by [Faily and Flechais 2011] and also persona campaigns, as suggested by [Matthews, Judge and Whittaker 2012], where updates concerning the personas are continually sent to the product developers. This feature may be important, as it appears that designers like continued access to the actual user data, aside from the persona description itself, to aid them in their decision-making Matthews. Moreover, the 2,491 user groups not currently reached by the organization is an interesting future research problem also that could be addressed with a gap analysis. An advantage of our system is that the data is archived with the system, readily available for more detailed analyses.

We are also investigating approaches to improve the core algorithm, such as tensor decomposition as an alternative to NMF that could provide means to systematically capture temporal dynamics, and we are experimenting with a wide range of topical classifiers. We used the LDA topics for characterizing videos content; however, there are other candidate features that we would like to explore, including the use of social media comments by others to classify the content [c.f., Kang and Lerman 2017]. Also, we are exploring additional methods of social comments, including facial recognition, as the demographic information for these accounts is not available. A careful selection of additional features could improve the performance of our model. Most importantly, we are planning to conduct an in-depth field evaluation, such as that reported in [Dharwada, Greenstein, Gramopadhye and Davis 2007] of the system with actual journalists in multiple roles in order to gauge their use and effectiveness within an organizational setting. Finally, we note that in addition to being a standalone method, we also see several advantages of our approach in conjunction with conventional off-line methods of persona creation (e.g., ethnography). Future research should aim at employing qualitative datasets along with quantitative ones.

8 CONCLUSION

In this research, we show that personas can be rapidly and automatically created from large scale, aggregated user data from major online social media platforms, resulting in personas that are based on behavioral data that reflects real people and created from

sizeable data quantities permitting quantitative analysis. We evaluated our persona generation methodology, showing that our method generates realistic and stable personas that have predictive ability. Although specifically focusing on digital content creators, our approach is flexible and resilient for application in a wide range of contexts where user data needs to be transformed into easy-to-understand representations for decision-making and customer insights. Automatic persona generation is an ongoing research agenda we are committed to carry out, improving the system and conducting user studies to show its practical benefits. As such, the current research should be seen as an opening to a host of related studies, in addition to being a major step forward in the persona creation research, paving the way for what we call “persona analytics”.

ACKNOWLEDGMENTS

We thank the many journalists at Al Jazeera News Media Network for their collaboration in this research.

REFERENCES

ABBAR, S., AN, J., KWAK, H., MESSAOUI, Y. AND BORGE-HOLTHOEFER, J. 2015. Consumers and Suppliers: Attention asymmetries. A Case Study of Aljazeera’s News Coverage and Comments. In *Proceedings of the Computation+Journalism Symposium 2015*, New York, NY, 2-3 October 2015.

ADLIN, T. AND PRUITT, J. 2010. *The Essential Persona Lifecycle: Your Guide to Building and Using Personas*. Morgan Kaufmann Publishers Inc.

ALCHEMY TAXONOMY API 2017. IBM.

AN, J., CHO, H., KWAK, H., HASSEN, M.Z. AND JANSEN, B.J. 2016. Towards Automatic Persona Generation Using Social Media. In *2016 IEEE 4th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, 206-211.

AN, J., KWAK, H. AND JANSEN, B.J. 2016. Validating Social Media Data for Automatic Persona Generation. In *The Second International Workshop on Online Social Networks Technologies (OSNT-2016), 13th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA2016)*, Agidar, Morocco.

AN, J., KWAK, H. AND JANSEN, B.J. 2017. Automatic Generation of Personas Using YouTube Social Media Data. In *Proceedings of the 50th International Conference on System Sciences (HICSS-50)*, Waikoloa, Hawaii, 833-842.

AN, J., KWAK, H. AND JANSEN, B.J. 2017. Personas for Content Creators via Decomposed Aggregate Audience Statistics. In *The 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2017)* IEEE, Sydney , Australia

- BEYER, H. AND HOLTZBLATT, K. 1998. *Contextual Design: Defining Customer-centered Systems*. Morgan Kaufmann Publishers Inc.
- BLEI, D.M., NG, A.Y. AND JORDAN, M.I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3, 993--1022.
- BLOMQUIST, Å. AND ARVOLA, M. 2002. Personas in Action: Ethnography in an Interaction Design Team. In *Proceedings of the Proceedings of the second Nordic conference on Human-computer interaction*, Aarhus, Denmark2002 ACM, 572044, 197-200.
- CASELLA, G. AND BERGER, R.L. 2001. *Statistical Inference*. Duxbury Press, Pacific Grove, CA.
- CHA, M., KWAK, H., RODRIGUEZ, P., AHN, Y.-Y. AND MOON, S. 2007. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *Proceedings of the 7th ACM SIGCOMM conference on Internet Measurement*, 1-14.
- CHAPMAN, C.N., LOVE, E., MILHAM, R.P., ELRIF, P. AND ALFORD, J.L. 2008. Quantitative evaluation of personas as information. In *Human Factors and Ergonomics Society 52nd Annual Meeting*, New York, NY, 1107-1111.
- CHAPMAN, C.N. AND MILHAM, R.P. 2006. The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. In *Human Factors and Ergonomics Society Annual Meeting*, San Francisco, CA, 634-636.
- CHEN, X., PANG, J. AND XUE, R. 2014. Constructing and Comparing User Mobility Profiles. *ACM Transactions on the Web (TWEB)* 8, Article 21.
- CLARKE, M.F. 2015. The Work of Mad Men that Makes the Methods of Math Men Work: Practically Occasioned Segment Design. In *Proceedings of the Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul, Republic of Korea2015 ACM, 2702493, 3275-3284.
- COOPER, A. 2004. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2nd Edition)*. Pearson Higher Education.
- DHARWADA, P., GREENSTEIN, J.S., GRAMOPADHYE, A.K. AND DAVIS, S.J. 2007. A Case Study on Use of Personas in Design and Development of an Audit Management System. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, Baltimore, Maryland, 469-473.
- DREGO, V.L. AND DORSEY, M. 2010. The ROI Of Personas Forrester Research.

DRUTSA, A., GUSEV, G. AND SERDYUKOV, P. 2017. Periodicity in User Engagement with a Search Engine and Its Application to Online Controlled Experiments. *ACM Trans. Web* 11, 1-35.

ERIKSSON, E., ARTMAN, H. AND SWARTLING, A. 2013. The Secret Life of a Persona: When the Personal Becomes Private. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Paris, France2013 ACM, 2481370, 2677-2686.

FAILY, S. AND FLECHAIS, I. 2011. Persona Cases: A Technique for Grounding Personas. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, BC, Canada2011 ACM, 1979274, 2267-2270.

FRIESS, E. 2012. Personas and Decision Making in the Design Process: An Ethnographic Case Study. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, Texas, USA2012 ACM, 2208572, 1209-1218.

GOODWIN, K. AND COOPER, A. 2009. *Designing for the Digital Age: How to Create Human-Centered Products and Services*. Wiley, Indianapolis, IN.

GRAY, R.M. 1984. Vector Quantization. *IEEE ASSP Magazine* 1, 4–29.

GRUDIN, J. AND PRUITT, J. 2002. Personas, participatory design and product development: An infrastructure for engagement. In *Participatory Design Conference*, 144-152.

GU!JÓNSDÓTTIR, R. AND LINDQUIST, S. 2008. Personas and Scenarios: Design Tool or a Communication Device. In *8th International Conference on Cooperative Systems (COOP'08)*, Carry-le-Rouet, France, 165-176.

GUO, F.Y., SHAMDASANI, S. AND RANDALL, B. 2011. Creating Effective Personas for Product Design: Insights from a Case Study. In *Internationalization, Design and Global Development: 4th International Conference, IDGD 2011, Held as part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011. Proceedings*, P.L.P. RAU Ed. Springer Berlin Heidelberg, Berlin, Heidelberg, 37-46.

JANSEN, B.J., AN, J., KWAK, H., HASSEN, M.Z. AND CHO, H.Y. 2016. Efforts Towards Automatically Generating Personas in Real-time Using Actual User Data. In *Proceedings of the Qatar Foundation Annual Research Conference 2016*, Doha, Qatar, 22-23 March 2016 QF, ICTPP3230.

JANSEN, B.J., SOBEL, K. AND COOK, G. 2011. Classifying Ecommerce Information Sharing Behaviour by Youths on Social Networking Sites. *Journal of Information Science* 37, 120-136.

- JOLLIFFE, I. 2002. *Principal component analysis*. John Wiley & Sons, Ltd., New York.
- JUDGE, T., MATTHEWS, T. AND WHITTAKER, S. 2012. Comparing Collaboration and Individual Personas for the Design and Evaluation of Collaboration Software. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, Texas, USA2012 ACM, 2208344, 1997-2000.
- JUNG, S., AN, J., KWAK, H., AHMAD, M., NIELSEN, L. AND JANSEN, B.J. 2017. Persona Generation from Aggregated Social Media Data. In *ACM Conference on Human Factors in Computing Systems 2017 (CHI2017)*, Denver, CO.
- KAHNEMAN, D. AND TVERSKY, A. 1972. Subjective probability: A judgment of representativeness. *Cognitive Psychology* 3, 430–454.
- KANG, J.-H. AND LERMAN, K. 2017. Effort Mediates Access to Information in Online Social Networks. *ACM Trans. Web* 11, 1-19.
- KRASHEN, S.D. 1984. Immersion: Why it works and what it has taught us. *Language and Society* 12, 61–64.
- KWAK, H. AND AN, J. 2014. Understanding news geography and major determinants of global news coverage of disasters. In *Proceedings of the Computation+Journalism Symposium 2014*, New York, NY, 24-25 October 2014.
- KWAK, H., AN, J. AND JANSEN, B.J. 2017. Automatic Generation of Personas Using YouTube Social Media Data. In *Hawaii International Conference on System Sciences (HICSS-50)*, Waikoloa, Hawaii, 833-842.
- LEE, D.D. AND SEUNG, S.H. 1999. Learning the Parts of Objects by Non-negative Matrix Factorization. *Nature* 401, 788-791.
- MAO, E. AND ZHANG, J. 2015. What Drives Consumers to Click on Social Media Ads? The Roles of Content, Media, and Individual Factors. In *2015 48th Hawaii International Conference on System Sciences*, 3405-3413.
- MARSDEN, N. AND HAAG, M. 2016. Stereotypes and Politics: Reflections on Personas. In *Proceedings of the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, Santa Clara, California, USA2016 ACM, 2858151, 4017-4031.
- MASSANARI, A.L. 2010. Designing for Imaginary Friends: Information Architecture, Personas, and the Politics of User-Centered Design. *New Media & Society* 12, 401-416.
- MATTHEWS, T., JUDGE, T. AND WHITTAKER, S. 2012. How Do Designers and User Experience Professionals Actually Perceive and Use Personas? In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Austin, Texas, USA2012 ACM, 2208573, 1219-1228.

- MCGINN, J. AND KOTAMRAJU, N. 2008. Data-driven Persona Development. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Florence, Italy2008 ACM, 1357292, 1521-1524.
- MCHUGH, M.L. 2012. Interrater reliability: the kappa statistic. *Biochemia Medica* 22, 276–282.
- MIASKIEWICZ, T., GRANT, S.J. AND KOZAR, K.A. 2009. A Preliminary Examination of Using Personas to Enhance User-Centered Design. In *AMCIS 2009 Proceedings*, Article 697.
- MULDER, S. AND YAAR, Z. 2006. *The User is Always Right: A Practical Guide to Creating and Using Personas for the Web*. New Rider, Berkely, CA.
- NIELSEN, L. 2004. Engaging Personas and Narrative Scenarios. In *Department of Informatics, Copenhagen Business School, Samfundslitteratur PhD Dissertation*.
- NIELSEN, L. AND HANSEN, K.S. 2014. Personas is Applicable: A Study on the Use of Personas in Denmark. In *Proceedings of the Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, Toronto, Ontario, Canada2014 ACM, 2557080, 1665-1674.
- PORTIGAL, S. 2008. Persona Non Grata.
- PRUITT, J. AND ADLIN, T. 2005. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufmann Publishers Inc.
- PRUITT, J. AND GRUDIN, J. 2003. Personas: Practice and Theory. In *Proceedings of the Proceedings of the 2003 conference on Designing for user experiences*, San Francisco, California2003 ACM, 997089, 1-15.
- REVELLA, A. 2015. *Buyer Personas: How to Gain Insight into your Customer's Expectations, Align your Marketing Strategies, and Win More Business*. Wiley.
- RODDEN, K., HUTCHINSON, H. AND FU, X. 2010. Measuring the User Experience on a Large Scale: User-centered Metrics for Web Applications. In *Proceedings of the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, Georgia, USA2010 ACM, 1753687, 2395-2398.
- RÖNKKÖ, K. 2005. An Empirical Study Demonstrating How Different Design Constraints, Project Organization and Contexts Limited the Utility of Personas. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 220a-220a.
- RÖNKKÖ, K., HELLMAN, M., KILANDER, B. AND DITTRICH, Y. 2004. Personas is not Applicable: Local Remedies Interpreted in a Wider Context,. In *Proceedings of the Proceedings of the eighth conference on Participatory design: Artful integration:*

interweaving media, materials and practices - Volume 1, Toronto, Ontario, Canada2004
ACM, 1011884, 112-120.

SHURADZE, G. AND WAGNER, H.T. 2016. Towards a Conceptualization of Data Analytics Capabilities. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 5052-5064.

SMITH, W.R. 1956. A Product Differentiation and Market Segmentation as Alternative Marketing Strategies. *Journal of Advertising* 21, 3-8.

STERN, B.B. 1994. A Revised Communication Model for Advertising: Multiple Dimensions of the Source, the Message, and the Recipient. *Journal of Advertising* 23, 5-15.

TESCH, R. 1990. *Qualitative Research: Analysis Types and Software Tools*. Psychology Press.

ZHANG, X., BROWN, H.-F. AND SHANKAR, A. 2016. Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, Santa Clara, California, USA2016 ACM, 2858523, 5350-5359.