Validating Social Media Data for Automatic Persona Generation

Jisun, An, Haewoon Kwak, Bernard J. Jansen
Qatar Computing Research Institute
Hamad bin Khalifa University
Doha, Qatar
jan@qf.org.qa, hkwak@qf.org.qa, jjansen@acm.org

***Abstract*— Using personas during interactive design has considerable potential for product and content development. Unfortunately, personas have typically been a fairly static technique. In this research, we validate an approach for creating personas in real time, based on analysis of actual social media data in an effort to automate the generation of personas. We validate that social media data can be implemented as an approach for automating generating personas in real time using actual YouTube social media data from a global media corporation that produces online digital content. Using the organization's YouTube channel, we collect demographic data, customer interactions, and topical interests, leveraging more than 188,000 subscriber profiles and more than 30 million user interactions. Then, we conduct statistical analysis on the social media data to determine whether the data could lead to the generation of valid personas based on statistically difference market segments. Findings show that customers can be segmented using product topics by gender and age based using social media data. However, our findings also show that the data is biased by the content created. The results offer insights into competitive marketing and product preferences for the consumers of the online digital content. Implications are that personas can be generated in real-time using social media data, instead of a time-consuming manual development process.**

Keywords*— *Persona; marketing; online news; design method; scenario; user-centered design.

I. INTRODUCTION

A persona is a concept used in marketing and advertising for abstract user representation, although the concept been adopted in a variety of other fields, such as system design [1]. Personas are representations of groups or segments of users that share common behavioral characteristics. Though representing a fragment of users, a persona is generally developed in the form of an explicit but fictitious individual, accompanied by a detailed narrative that represents the collection of users possessing similar behaviors or characteristics.

The persona narrative can contain a variety of both demographic and behavioral details about socio-economic status, gender, hobbies, family members, friends, possessions, among many other data in order to make the fictitious individual appear as real person to the content or product developers,. Similarly, the narrative of a persona can addresses the goals, needs, wants, frustrations, and other emotional aspects of the fictitious individual that are pertinent to the content or product being designed. The fictitious individual represented in the persona is generally given a name and an image to assist the content marketers in focusing on the particular user segment [2].

II. RELATED WORK

From systems to services to advertisements, personas are beneficial for administering design decisions and evaluating product ideas. Within the domain of content creation, personas are considered and have been shown to be a worthwhile technique in directing creators and designers. One creates a persona for clarifying and synthesizing descriptions of user sections, with the idea being that a persona assists a creator in focusing on the behavior patterns, wants, and needs of the segment of users. Products, especially those deployed to market, can have multiple user segments that each require a persona, so it can be complex during the development of the granularity of data and execution of data integration in a meaningful way.

Preferably, the construction of a persona is based on actual consumer data research of a product's intended user base(s). Typically, this has been a foremost issue, as user data from the field has generally been gathered via surveys, focus groups, or ethnography methodologies. Unfortunately, a common problem with creating personas via these manual methods is that they are many times not based on first-hand user data or the data set is not of a sample size that can be considered statistically significant. Additionally, utilizing these processes for persona development can be costly and time-consuming.

Therefore, on the ground, persona development does not actually come from user studies or genuine current user behavioral data; instead, the personas are based on the assumptions, experiences, or expectations of executives, marketers, or designers or the personas quickly become out of date. This process results in personas that are not believable or do not actually represent the real current or targeted users [3, 4]. This problem is exacerbated in actual commercial products in fast moving and competitive market areas, where the user base is multivariate and nearly always opens to the possibility of flux.

Currently, there are many unanswered questions concerning the automatic generate of personas. *Can the procedure of actual user behavioral data be used for personas development? Can online data deliver the rich demographic insights common in personas? Can personas be established in near real time? Can personas be frequent updated?* These are some of the questions that motivate our research.

III. RESEARCH OBJECTIVE

In the research presented here, we put forward the premise that real user behavior and related demographic data concerning users of a product, service, or content [5] can be rapidly and inexpensively collected from variety of social platforms and analyzed in order to generate personas in real-time. Attaining such an objective means that personas based on social media data (a) represent the current users of the product and (b) the social media data can distinguish among different market segments in a statistically meaningful manner. This is the research objective that we investigate, awe present the current state of research and system development here.

IV. DATA COLLECTION AND RESEARCH DESIGN

Using actual user data from AJ+, an online media and mobile outlet, we validate our premise concerning social media data. In the news industry, audience inclinations have been rather ignored by journalists mainly because of the lack of accurate measurements before the era of online news. Several studies point out large differences between news production and consumption patterns by using the number of articles on particular topics and their views in news websites [e.g., 6]. Precise understanding of audiences becomes more important to attract audience and increase the consumption of digital content.

Consequently, we deem AJ+ an excellent organization for both data collection, and our research with real AJ+ user data shows the value of automatic persona generation for news readership research in particular, although we consider the results transferable to other industry segments, using social media data.

A. Data Collection Organization

An online news channel from Al Jazeera Media Network, AJ+ (http://ajplus.net/) that is natively digital with a presence only on social media platforms and on a mobile app. Its media concept is unique in that AJ+ was designed from the ground up to service news in the medium of viewer, with no redirect to a website. Although with a presence on iOS and Android apps, AJ+ is based primarily on social platforms. Thus, the digital content is specifically designed to be viewed in the Facebook newsfeed, YouTube Channel, or Twitter Timeline for the audience segments that are most active on those platforms.

In its storytelling formats, app design, and video development, AJ+ has been innovative in its experimentation, receiving significant press [7]. During the period of this study, AJ+ was the second largest producer of video on Facebook, had more than 3 million Facebook followers, 195,000 Twitter followers, and 188,000 YouTube subscribers. AJ+'s engagement rate at this time was 600% (i.e., their news products are engaged by 6x their base) [8].

AJ+ has a critical need for automatic and real time generation of personas to guide digital content, media, and system planning and design, given the prerequisite for rapid and media specific development of content in a competitive and fluid information market.

Therefore, in pursuit of our overall research objective of validating social media data for automatically generating personas in real time, for the research reported in this manuscript, we are specifically interested in understanding the AJ+ audience by identifying (1) whom are they reaching (i.e., market segments) and (2) is there a significant difference in these market segments that is reflected in the social media data.

From this, combined with other user data [e.g., 9, 10], we can validate the design a system using actual user data for personas generation in near real time.

B. Data Collection (YouTube)

For data collection, we focus on the YouTube channel in the research reported here, although could and are planning on doing similar analysis for Twitter and Facebook. The chief reason to focus on YouTube is that the analytics platform gives the most detailed statistics for every video, compared to other channels. An example of an AJ+ YouTube video is shown in Figure 1, noting the likes/dislikes, shares, and views.

Fig. 1. An example of YouTube Video from the AJ+ YouTube Channel.

Being quite robust, the YouTube analytics platform provides, for each of AJ+ videos, user profile data (e.g., gender, age, country location, and which site the user comes from) at an aggregate level. We use these datasets to validate if social media data can provide the demographic information needed for and automatic persona generation system.

One can access the statistics in the YouTube analytics platform by the YouTube APIs [1]. Table I displays the parameters we used for calling the APIs. Even though there are other metrics besides viewCount (the number of views) and viewerPercentage, such as likes or comments, the sheer value of the number of likes or comments is much lower than that of views. Therefore, we focus only on viewCount and viewerPercentage. As a note, the data from a YouTube

[1] https://developers.google.com/youtube/analytics/

channel is private and available only to an owner of the YouTube channel (i.e., AJ+ in this case), and thus not publicly accessible.

TABLE I. QUERY PARAMETERS AND METRICS USED FOR YOUTUBE APIS.

Dimension	Metric
ageGroup	viewerPercentage
gender	viewCount
country	
month	
day	

V. RESULTS

A. Exploratory Analysis

First, we present some overall statistics from the AJ+ YouTube channel page.

During the study, there were 2807 AJ+ videos posted to the AJ+ YouTube Channel. These 2807 AJ+ videos had more than 30 million views by users from 217 countries.

Overall, the AJ+ audience is worldwide, with the topmost three countries, in terms of viewership, being Canada, Great Britain, and the United States (US), with each representing 2.44 percent of total viewership in terms of the number of unique videos watched. Concerning the total number of views, the US is the largest country market segment, with about 49.4 percent of video views coming from the US. It is also interesting to note that, although AJ+ was designed to target the US market, a majority of viewers come from outside the US, challenging to have a comprehensive understanding about their viewers.

Concerning gender and age distribution, 20.9 percent of viewers were female, with 79.1 percent being male. YouTube views are classified into multiple age categories (13-17 years, 18-24 years, 25-34 years, 35-44 years, 45-54 years, 55-64 years, and 65 years and older). The age percentages are shown in Table II. As AJ+ is designed to target young generation by adopting social platforms, it is expected that young adult male is the biggest segment.

Some videos showed a worldwide appeal, with 100 videos being viewed in 100 or more countries. Conversely, there were about 100 videos that were viewed by users from 5 or fewer countries. In terms of the actual number of views, the viewership counts per individual videos follow a power law distribution, with a small number of videos being viewed a lot and a large number of videos being viewed a small number of times. Such skewed popularity of videos is one of the well-known characteristics in YouTube [8].

TABLE II. PERCENTAGE OF VIDEO VIEWINGS BY AGE GROUPINGS.

Age Grouping	Percentage
13-17	4.46%
18-24	38.04%
25-34	24.80%
35-44	16.39%
45-54	8.55%
55-64	4.50%
65+	3.26%
	100.00%

B. Research Objective Analysis

For more general content consumption patterns, we transform video-level viewing patterns into topic-level ones. For this, we classified the topic of each AJ+ video by using its title and short description (normally a few sentences). We used the Alchemy Taxonomy API[2] for topic classification. Alchemy Taxonomy API supports over 1,000 topics, and these are organized as a hierarchical structure. Based on our analysis, AJ+ videos are classified into 22 tier-1 level topics.

Moving to our specific research objective, we first explore the relationship between the content produced by the company and the content consumed by viewer, in the aggregate. This relationship is shown in Table III, with topic of videos, percentage of videos produced in that topic, the ranking of topic by videos produced, the percentage of videos viewed within that topic, and the ranking of topics by videos viewed.

TABLE III. VIDEO TOPICS WITH PERCENTAGE OF VIDEOS PRODUCED AND VIEWED IN THAT TOPIC AND RANK OF EACH.

Topic	Percentage of Videos Produced	Rank by Videos Produced	Percentage of Video Views	Rank by Video Views
art and entertainment	9.8427%	3	7.9433%	4
automotive and vehicles	0.6596%	21	0.2828%	21
business and industrial	4.4140%	7	2.3198%	11
careers	0.1522%	24	0.0572%	24
education	2.1816%	13	1.8714%	14
family and parenting	3.3486%	9	2.2935%	12
finance	0.9132%	16	0.6396%	17
food and drink	2.4860%	12	2.4445%	10
health and fitness	2.9934%	11	3.0929%	8
hobbies and interests	0.8625%	17	7.5303%	5
home and garden	0.8118%	18	0.3608%	20
law, govt and politics	27.9554%	1	25.5790%	1
news	0.8118%	19	0.7451%	15
not defined	0.2029%	23	0.1870%	22

[2] http://www.alchemyapi.com/products/alchemylanguage/taxonomy

Topic	Percentage of Videos Produced	Rank by Videos Produced	Percentage of Video Views	Rank by Video Views
pets	0.8118%	20	0.6623%	16
real estate	0.3044%	22	0.1699%	23
religion and spirituality	4.7184%	6	6.0426%	6
science	3.2978%	10	3.4141%	7
shopping	0.9640%	15	0.4708%	19
society	13.0391%	2	17.5134%	2
sports	4.3125%	8	2.6744%	9
style and fashion	1.1669%	14	0.6062%	18
technology and computing	5.9361%	5	2.2212%	13
travel	7.8133%	4	10.8778%	3
	100%		100%	

From Table III, two findings are apparent upon analysis. First, of the 23 video topical classifications, only 7 of the 23 topics have the same ranking in both videos produced by topic and videos viewed by topic, with 17 of the topics having different rankings based on the two classifications. Second, although the rankings are not identical, the rankings are also somewhat similar. For example, the topic *travel* is ranked 4th by videos produced and 3rd by videos viewed. So, while not identical, there is some apparent correlation among the two, as one would expect (i.e., a consumer can't view a video unless it is produced). In fact, a Spearman's rho shows a 0.85 correlation between the two rankings.

With the aggregate analysis completed, we then examined gender and age. The results of the gender analysis are shown in Table IV, with topical classification, ranking of topics by videos viewed by females, and ranking of topics by videos viewed by males.

TABLE IV. VIDEO TOPICS AND RANKING BY VIDEO VIEWING WITHIN THAT TOPIC BY MALE AND FEMALE VIEWERS OF THE AJ+ CHANNEL.

Topic	Rank of Topic by Female Viewing	Rank of Topic by Male Viewing
art and entertainment	5	4
automotive and vehicles	21	21
business and industrial	14	10
careers	24	24
education	13	14
family and parenting	10	13
finance	17	17
food and drink	8	11
health and fitness	9	8
hobbies and interests	4	5
home and garden	20	20
law, govt and politics	1	1
news	16	15
not defined	22	22
pets	18	16
real estate	23	23
religion and spirituality	6	6
science	7	7
shopping	19	19
society	2	2
sports	11	9
style and fashion	15	18
technology and computing	12	12
travel	3	3

As shown in Table IV, the rankings of video viewing by males and females are generally similar, with 12 of the 24 topics having the same ranking for both genders. The top seven most viewed topics being the same for females and males, in order, are *law, govt and politics, society, travel, hobbies and interests, art and entertainment, religion and spirituality, and science*. So, the rankings appear very similar overall, and a Spearman's rho shows only a 0.97 correlation between the two rankings.

However, there are some differences beyond the 'hit' topics, which one can use to discriminate between genders. Several Internet marketing companies, such as OpenSlate[3], have reported the differences in the most popular contents for men and women. Beauty & style is dominant for women (62%) and Gaming is for men (51%). The rankings in Table IV are consistent with such previous report; in the table, the most popular category for female is style & fashion, and that for male is hobbies and interests. This means that, even in a single news channel, the gender differences emerge, and thus, understanding the audience correctly is important.

We then examined topical ranking by age grouping, as shown in Table V, with Spearman's rho values displayed in Table VI. From Tables V and VI, there are several takeaways. First, there are some topics that rank identically (e.g., *careers, food and drink, law, govt and politics*, *society*, *religion and spirituality*, and *travel*) or nearly identical (e.g., *art and entertainment*, *automotive and vehicles*, *education*, *home and garden*, and *real estate*) for almost all age groupings. This is nearly half of topical categories.

[3] https://www.openslatedata.com/news/how-content-powers-share-shift/

TABLE V. VIDEO TOPICS AND RANKING BY VIDEO VIEWING WITHIN THAT TOPIC BY AGE GROUPING VIEWERS OF THE AJ+ CHANNEL.

Topic	Rank of 13-17 Age Group	Rank of 18-24 Age Group	Rank of 25-34 Age Group	Rank of 35-44 Age Group	Rank of 45-54 Age Group	Rank of 55-64 Age Group	Rank of 65+ Age Group
art and entertainment	3	3	3	3	4	4	4
automotive and vehicles	21	21	21	20	21	20	20
business and industrial	7	6	6	9	10	9	8
careers	24	24	24	24	24	24	24
education	13	13	13	13	13	14	13
family and parenting	9	10	10	8	7	8	7
finance	17	16	16	17	18	16	17
food and drink	12	12	12	12	12	12	12
health and fitness	11	9	11	11	11	10	11
hobbies and interests	15	14	14	14	14	13	14
home and garden	20	20	20	21	20	21	21
law, govt and politics	1	1	1	1	1	1	1
news	19	17	17	15	15	15	15
not defined	23	23	22	22	22	22	23
pets	18	19	19	19	19	17	19
real estate	22	22	23	23	23	23	22
religion and spirituality	5	5	5	5	5	5	5
science	10	11	8	6	6	6	6
shopping	16	18	18	18	17	19	18
society	2	2	2	2	2	2	2
sports	6	7	7	7	9	11	10
style and fashion	14	15	15	16	16	18	16
technology and computing	8	8	9	10	8	7	9
travel	4	4	4	4	3	3	3

TABLE VI. SPEARMAN RHO VALUES BY AGE GROUPING OF AJ+ YOUTUBE CHANNEL.

	13-17	18-24	25-34	35-44	45-54	55-64	65+
13-17	1.0000	0.9913	0.9904	0.9757	0.9713	**0.9557**	0.9704
18-24	0.9913	1.0000	0.9930	0.9757	**0.9687**	**0.9643**	0.9730
25-34	0.9904	0.9930	1.0000	0.9887	0.9800	0.9730	0.9835
35-44	0.9757	0.9757	0.9887	1.0000	0.9930	0.9826	0.9930
45-54	0.9713	**0.9687**	0.9800	0.9930	1.0000	0.9878	0.9948
55-64	**0.9557**	**0.9643**	0.9730	0.9826	0.9878	1.0000	0.9904
65+	0.9704	0.9730	0.9835	0.9930	0.9948	0.9904	1.0000

For the other categories, the distribution among the age groupings is more distributed, but even then, there is generally not too much variance, although there are some topics that are exceptions (e.g., *business and industrial*, *family and parenting*, *science*, and *sports*).

This observation is supported by Spearman's rho analysis values shown in in Table VI. The rankings within each age grouping for video topics are highly correlated, overall. However, as noted in the bolded values, there is some age divergence, supporting the qualitative observations that certain topics, a handful, vary among age groups.

VI. Discussion and Future Direction

Our research shows that social media data or the automatic persona generation system is quite robust in identifying market showing that one can use actual user behavior that is rapidly collected and analyzed in order to generate personas in real-time. However, there are several further research and development fronts that we are pursuing.

A. Rich Persona Attributes Beyond Demographic Data

The main strength of our approach is that we benefit from actual user data, reducing time and cost for generating personas relative to traditional methods, and thus our approach is suitable for real-time persona generations. With our technique, we do not need to cautiously sample users for interviews to develop personas; instead, we can analyze and extract representative personas from millions of users by using publicly available online social media data.

Also, our research is an initial point for creating personas for a vast number of other applications and services without too much manual efforts. If we can leverage more rich information of a user that is usually considered for defining personas, such as ethnicity, and socio-economic status, our methods and results would become more and more useful. We are exploring these methods for future work.

One possibility is to extract demographic information from shared links on Facebook [9] or from Twitter bios or from Google+ accounts. Shared links could reveal things like economic status. Prior research has shown that affluent customers (high-end luxury product websites) and budget conscious customer (price aggregation or discount websites) can be distinguished by websites they visited [10]. Using these as examples, we are investigating the extraction of rich information from shared links and websites.

B. Scalability of a Persona System Using Social Media

One possible concern of our approach is the scalability of a system relying on social media data. We do not view this as a major concern, as the API requires only a few number of HTTP requests for each user, and we only need 'newer data' once the initially collected. This implies that the number of required requests does not increase radically over time.

C. Benefits for Journalists of Using Social Media

Using social media for persona generation provides initial demographic information and more rich behavioral details. Our development approach is derived from collaboration with journalists who actually benefit from the generated personas. Journalists wish to have the realistic view of the actual users so that they can reach readers with better titles, content, and article framing. In this sense, our persona research offers a strong foundation, as the resulting personas describe what topics readers are interested in.

VII. conclusion

In this research, we show that social media data can be used to create real time and automatic persona generation. We have taken the initial fruitful steps to move personas creation from a manual, time intensive process to that which is automated and in near real time. We are continuing systems development to enhance features for personas selection and persona filtering. We are currently investigating using other data sources [11-13], including additional social media platforms as well as off-line sources to provide richer demographic attributes, attitudinal characters, and other aspects for rounding out the generated personas.

References

[1] A. Cooper, *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. US: Sam, 1999.

[2] T. Matthews, T. Judge, and S. Whittaker, "How Do Designers and User Experience Professionals Actually Perceive and Use Personas?," in *CHI2012*, 2012, pp. 1219–1228.

[3] T. Judge, T. Matthews, and S. Whittaker, "Comparing collaboration and individual personas for the design and evaluation of collaboration software," in *CHI.2012*, 2012, pp. 1997-2000.

[4] J. E. Nieters, S. Ivaturi, and I. Ahmed, "Making Personas Memorable," in *CHI 2007*, 2007, pp. 1817–1824.

[5] B. J. Jansen, K. Sobel, and G. Cook, "Classifying Ecommerce Information Sharing Behaviour by Youths on Social Networking Sites," *Journal of Information Science,* vol. 37, pp. 120-136, 2011.

[6] S. Abbar, J. An, H. Kwak, Y. Messaoui, and J. Borge-Holthoefer, "Consumers and Suppliers: Attention asymmetries. A Case Study of Al Jazeera's News Coverage and Comments," in *Computational Journalism Symposium*, 2015.

[7] J. Roettgers. (2015, 30 July) How Al Jazeera's AJ+ Became One of the Biggest Video Publishers on Facebook. *Variety*. Available: http://goo.gl/76DZBP

[8] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "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*, 2007, pp. 1–14.

[9] J. An, H. Y. Cho, H. Kwak, M. Z. Hassen, and B. J. Jansen, "Towards Automatic Persona Generation Using Social Media," in *The Third International Symposium on Social Networks Analysis, Management and Security (SNAMS 2016)*, 2016, pp. 1-6.

[10] A. Odlyzko, "Privacy, economics, and price discrimination on the internet," in *Proceedings of the 5th international conference on Electronic commerce*, 2003, pp. 355–366.

[11] D. Martín-Albo, L. A. Leiva, J. Huang, and R. Plamondon, "Strokes of insight: User intent detection and kinematic compression of mouse cursor trails," *Information Processing & Management,* vol. 52, pp. 989-1003, 2016.

[12] J. He, P. Qvarfordt, M. Halvey, and G. Golovchinsky, "Beyond actions: Exploring the discovery of tactics from user logs," *Information Processing & Management,* vol. 52, pp. 1200-1226, 2016.

[13] B. J. Jansen and S. Simone, "Bidding on the Buying Funnel for Sponsored Search Campaigns," *Journal of Electronic Commerce Research,* vol. 12, pp. 1-18, 2011.