Peeking Into Minds of iGeneration via Lyrics

of Most Popular Songs over 50 Years

Completed Research

Yong Seog Kim
Utah State University
yong.kim@usu.edu

Abstract

In this paper, we identify unique textual patterns from popular songs in iGeneration in which people feel comfortable with the Internet and mobile technology to make contact with friends and family through social media sites. To this end, we calibrate data-driven classifiers using terms and their frequencies in the lyrics of Billboard’s Year-End Hot 100 songs between 1965 and 2015. We find that iGeneration shows several contrasting characteristics in terms of long lyrics and frequent emotional terms related to aspiration, self-focus, negation, and curse.

Keywords

iGeneration, text mining, word cloud, Fourth Turning theory, Millennial saeculum.

Introduction

In this paper, we intend to identify unique textual patterns from popular songs in iGeneration (or Generation Z) in which people feel comfortable with the Internet and mobile technology to make contact with friends and family through social media sites. This study is in part motivated by the Fourth Turning theory (Strauss and Howe 1992), which states that a set of four generational archetypes with new social, political, and economic climates have recurred about every 80-90 years called “saeculum” in American history. In particular, as the last archetype of the Millennial saeculum, iGeneration may create a tumultuous political and societal environment different from those of three prior generations, the Baby Boom generation followed by Generations X and Y. Therefore, it is important to understand what concerned people in iGeneration equipped with information and communication technologies, and how they expressed their hopes, ideals, anger, and frustrations.

This study also intends to integrate various studies in sociology, music, history, psychology, information system, and data analytics. That is, while many studies shown in the following literature review have studied the characteristics of generations in terms of work ethic, societal relationships, political views, racial perspectives and adoption of new information technologies, there are very few studies that use data-driven approach to quantitatively validate differences among generations reported in these studies. To this end, we briefly review possible impacts of songs on society and culture, and describe the demographic characteristics of Generations X, Y, and iGeneration. Then we calibrate data-driven classifiers using terms and their frequencies extracted from the lyrics of 5,100 songs over 50 years as inputs to identify unique emotional and psychological terms in each generation. Finally, we conclude our paper by suggesting several future research directions.

Literature Review on Songs, Society and Generations

Songs and Society

According to the celebrated Chinese philosopher Confucius, “music produces a kind of pleasure which human nature cannot do without.” To pursue of such emotional pleasure that music can associate, people in every society consume it in their daily lives and in various situations, resulting in US$15.7 billion music business in 2016 (IFPI Global Music Report 2017). In USA, an average American listens to 4.5 hours of music each day, which leads to approximately 2 million songs in his or her lifetime (The Nielsen Company
Therefore, music (and hence musicians) can perceptually, emotionally and morally influence individuals and our society. For example, several research showed that musical features could stimulate psychophysical reactions (Bernardi et al. 2009; Kreutz et al. 2008) and alter our decision making processes (Marin and Bhattacharya 2010). Other studies also demonstrated that unexpected change in acoustic features such as pitch, loudness, or tempo, and intervals and regularities between the different music elements induced a strong emotional and physiological response in listeners (Arjmand et al. 2017; Coutinho and Cangelosi 2011; Juslin and Sloboda 2010; Altenmüller et al. 2002). In particular, musical structure and acoustic features seem to be more important in determining emotional reactions than the listener’s mood, affective involvement, personality or contextual factors (Scherer et al. 2002; North 2012; Reiner et al. 2010).

Music can also impact morals and behavior of listeners, especially of the youth and adolescents who are still susceptible to the ethics and regularities around them. For example, adolescents listening to many songs with lyrics that consist of terms such as sex, drugs and violence are likely to construct self-images that glorify or justify these types of behaviors (Johnson et al. 1995; Martinez 1997). At the same time, various music genres such as punk rock, heavy metal, and rap provide fans and audiences with distinct styles of dances, dress, slang, and gestures (Mauch et al. 2015). In particular, popular star artists such as Elvis Presley and Britney Spears with their unique styles (e.g., Elvis’ ducktail haircut and Britney’s bare midriff) greatly impact on their fans’ morals and attitudes (Walters and Spitzer 2003). For the same reason, almost all countries designate national anthems to encourage patriotic morals of their people (e.g., songs like “Yankee Doodle” and the “Star Spangled Banner” for American people).

Considering the fact that music can perceptually, emotionally, and morally influence individual listeners who are the main building blocks of a society, musical work ultimately influences the culture of our society, too (Walters and Spitzer 2003; Askin and Mauskapf 2014). This is why historians often investigate the distinct characteristics of a certain generation in the tunes, lyrics and sound of that time, believing that such characteristics reveal people’s reflection on what concerned them, how they saw issues, and how they expressed their hopes, ideals, anger, and frustrations (Walters and Spitzer 2003).

**Baby Boom Generation and Generations X, Y and Z**

While various authors have delimited the Baby Boom period differently, baby boomers are typically defined those born between mid-1940s and 1964, a period associated with a noticeable increase in the birth rate. As a group, baby boomers were the wealthiest with peak levels of income and abundant levels of food, apparel, and retirement programs (Jones 1980). However, boomers suffered dramatic social change in the United States due to the conflicts between the proponents of change and the more conservative individuals on issues such the Vietnam War, the civil rights movement, and the women’s rights (Jones 1980). The baby boomers used transistor radios to listen to rock and roll music and became the first generation to grow up with the television.

Many researchers and demographers define Generation X as adults born between 1965 and 1980 (e.g., Pew Research Center). According to demographers William Strauss and Neil Howe (1993), Generation Xers experienced a cultural shift from emphasizing the societal value of parental engagement with their children to encouraging the societal value of parental (in particular for maternal) self-actualization, resulting in sharp increase in divorce rates starting from the mid-1960s before it peaked in 1980. In particular, middle and upper class children mostly suffered from lacked adult supervision and staying at home alone mainly due to increased maternal participation for economic activities (Blakemore 2015). Therefore, Generation Xers became more peer-oriented and enjoyed grunge music genre whose lyrics often address themes such as social alienation, despair and apathy through topics on homelessness, suicide, rape, broken homes, drug addiction, self-loathing, and domestic abuse. Generation X was also known to be the first cohort to come of age with MTV and music videos.

Millennials (or Generation Y) are the generational cohort born between early 1980s and mid-1990s to early 2000s. Millennials were the first to grow up with computers in their home, which make them feel comfortable with using social media and digital technologies not only to remain connected with friends but also directly purchasing new products from factories and exchanging pre-owned products through the Internet. However, with a median household income of $40,581, Millennials earn 20% less than the generation before them due to the Great Recession and have benefited the least from the economic recovery following the Great Recession (Strauss and Howe 2000). To be competitive, Millennials are more likely to
have gone to university, but end up being in massive student debt due to skyrocketed cost of education. All these circumstances led them to have a skeptical view on long-term economic and social improvement, and less focus on forming new families, spouses or partners and children (Anderson 2013). Because of this characteristic, Millennials are often referred to Peter Pan generation who is most likely to stay with their parents for longer periods due to the high cost of housing and the relative affluence of older generations (Shaputis 2004; Palmer 2007). However, Millennials are known to pursue work-life balance through exercising more, eating smarter, and smoking less under far more diverse environments (Hershatter and Epstein, 2010; US Census Bureau 2015).

The last generation we consider in this study is Generation Z, people born from 2000s. Generation Z is the first generation to have been born after the popularization of the Internet, and hence is named as iGeneration. Naturally, Generation Z feel comfortable with the Internet and mobile technology and make contact with friends and family, and to develop new relationships through social media sites (Turner 2015). However, their heavy social media usage also reflect their loneliness, anxiety, and fragility caused from both September 11 terrorist attacks and the Great Recession in 2008. Generation Z also suffers from unsettlement and insecurity from a growing income gap and a shrinking middle-class in their era (Dupont 2015). These major societal and economic events not only teach them Generation Z to be independent with an entrepreneurial desire but also encourage conservative behaviors such as lower teen pregnancy rates, less drug and alcohol abuse, and higher on-time high school graduation rates compared with Millennials. Interestingly, this generation enjoys consuming pop, rock, hip-hop, country subgenres and new music trends such as K-Pop, trap music and bedroom pop.

Data Source and Data Engineering

This study investigates the lyrics of songs through text mining methods to diagnose the cultural spectrums of modern generations based on the very understanding that culture and music co-develop together (Balkwill and Thompson 1999; Balkwill et al. 2004). Ultimately, our study intends to provide insights about which ethical, emotional, and moral directions we should pursue to better cultivate our culture and society. In particular, we investigate the appearance of terms in the lyrics of songs over 50 years to relate key terms in songs to possible societal, political and economic changes in three generations, Generation X, Y, and Z. While, our study is similar to the project named Money, Love And Sex (Lamm 2013) in that both studies track the popularity of certain key terms from the lyrics of song, our study is unique in a sense that it analyzes the frequency trends of key terms as a means of understanding societal and cultural characteristics of each generation (e.g., iGeneration’s heavy dependence on the Internet and social media). In addition, our study will train classifiers to determine which generation each song belongs to only using extracted textual terms and their frequencies. Using textual terms and their frequencies, it may be possible to train classifiers to predict whether or not new songs can be placed within top charts (Interiano et al. 2018).

![Figure 1. Sample Datasets](https://www.kaggle.com/rakannimer/billboard-lyrics)

We first downloaded the initial dataset from Kaggle (https://www.kaggle.com/rakannimer/billboard-lyrics). This dataset consists of six variables (e.g., Rank at Year-End, Song title, Artist, Year, Lyrics, and Source) for each of Billboard’s Year-End Hot 100 songs between 1965 and 2015. When songs are
Peeking Into Minds of iGeneration via Lyrics

instrumental type with no voice or lyrics, their lyrics are simply denoted as “instrumental”. For our analysis, all of the 5100 songs with two engineered variables are used for analysis. The two engineered variables are Decades and ThreePeriods, which were all derived from Year feature. They were engineered so that we can investigate whether or not textual information extracted from lyrics can have a distinctive power to correctly classify the time period that songs were on the market. The Decades (Y1960s to Y2110s) are easily derived from Year feature, but the values of the ThreePeriods feature are subjectively engineered so that each value corresponds to one of three generations, Generation X, Y, and Z. That is, while the first period, Period1, is defined as 1965 – 1980 (which roughly corresponds to the time period of Generation X), the second and third periods are defined as 1981-2000 (i.e., the period of Generation Y) and 2001-2015 (i.e., the period of Generation Z), respectively. We show the snapshot of datasets with engineered features in Figure 1.

However, textual information in the lyrics of the original datasets are unstructured and hence cannot be processed by many popular classifiers such as Decision Tree (DT) model, an Artificial Neural Net (ANN) model, and a Naïve Bayes (NB) model. Therefore, it is necessary to process textual information in lyrics and transform them into a structured format. We use two operations, term extraction and term frequency transformation, available in Microsoft SQL Server 2016. The first operation, term extraction, is to generate the list of terms from the lyrics of all songs. Note that the lyrics of each song is represented by the terms it contains and their occurrences. To extract important terms from lyrics, we consider both TF(t) metric, that considers a term t as more relevant to semantics when it occurs more frequently within a lyric and IDF(t) metric that considers a term t as more discriminative when it occurs less frequently among lyrics. In particular, we use a combined metric of these two metrics, TF(Term Frequency) * IDF(Inverse document frequency), to include important and discriminative terms, resulting in a total of 4,365 terms. Note that however, it is possible that important information may be lost by ignoring word order, sentence structure, and context. Finally, we calculate the frequency of all 4,365 terms for each 5100 songs, resulting in the structured term frequency table with 169,979 records.

Validating Discriminative Power of Terms in Lyrics

We calibrate DT, ANN, and NB models because they are most widely applied for various research fields (DeTienne and DeTienne 2017; Kim et al. 2005; Kim and Moon 2012). Most of all, each of these algorithms has been well known for its speed and interpretation (DT), its robustness and accuracy (ANN), and its scalability (NB). Note that, in this study, it is not our main objective to train the most accurate model to classify which period (or decade) each song belongs to. Instead, we aim to test our concept that terms and their frequencies in lyrics have a discriminative power to profile the characteristics of songs in each time period (or decade). For readers who are interested in modeling the best predictive models may apply meta-classifiers (i.e., ensemble models such as random forest, gradient boosted forest) at the cost of interpretability and computation complexity.

To define sensitivity metrics, it is necessary to simplify our multi-class problem into a bi-class classification problem, where prediction models are trained to identify songs in “Period1” (denoted as positive records) and all the remaining songs (denoted as negative records). When models return their predictions for records in a test data, four possible cases are tabulated: the number of true positive (TP) cases when a prediction model correctly classifies a positive record as positive, and the number of false positive (FP) cases when it misclassifies a negative record as positive, the number of true positive (TN) cases when it a prediction model correctly classifies a negative record as negative, and the number of false negative (FN) cases when it misclassifies a positive record as negative. Then the sensitivity is defined as TP / (TP + FN) while accuracy is defined as (TP + TN) / (TP + TN + FP + FN).

For sensitivity analysis, a lift chart is used to visualize the predictive power of a prediction model. In a lift chart, y-axis represents sensitivity (or true positive rate or hit rate) for a corresponding x-axis value, representing a case when top x% of population in terms of the estimated probability of being positive are considered for classification. By definition, a lift chart for a random model is always a diagonal line because it will statistically correctly identify x% of positive records when top x% of population are classified. In this section, we are interested in three sensitivity analyses estimate a model’s ability to correctly detect songs in Period1 (i.e., Generation X), Period2 (i.e., Generation Y), and Period3 (i.e., Generation Z). However, due to the limited space, we only present two outcomes for Period1 and Period3.
The lift charts in Figure 2-a represent outcomes when classifiers try to identify songs in Period1 (i.e., 1965-1980) (i.e., songs with “Period1” class value are regarded as positive while others are negative). In Figure 2, a purple, a green, and a yellow line represents the lift chart of DT, ANN, and NB models, respectively. A red and a blue line represents the lift chart of an ideal and a random model, respectively. According to Figure 2-a, when top 43% of songs are chosen for a prediction task, the ideal model and a random model identifies 100% and 43% of positive samples, respectively, while ANN identifies 64% of positive samples and DT and NB identify about 47% of positive samples. However, when top 20% of songs are chosen for a prediction task, while ANN and DT identifies 37% and 29% of positive samples, a random model identifies only 20% of positive samples. Therefore, the lift (or improvement) of ANN and DT over a random model is calculated as 1.85 (=37%/20%) and 1.45 (=29%/20%), respectively. When top 10% of songs are chosen for a prediction task, the lift of ANN and DT over a random model results in 1.98 (=19.8%/10%) and 1.7 (=17%/10%). Overall, while all three classifiers perform much better than a random model, ANN performs best followed by DT and NB in the task of identifying songs in Period1.

Figure 2-a. Lift Charts of Period1 Class  Figure 2-b. Lift Charts of Period3 Class

In Figure 2-b, we present the lift charts of ANN, DT, NB, an ideal model, and a random model for a classification task to identify songs in Period3 (i.e., songs in Period3 are regarded as positive while others negative). According to Figure 2-b, when the top 20% of songs are chosen for a prediction task, ANN identifies about 48% of positive samples but DT and NB identifies 46% of positive samples, while an ideal and a random model identifies 70% and 20% of positive samples, respectively. Overall, ANN performs slightly better than DT and NB over a range of small sample sizes (between 0% and 23% of songs). However, more than 20% of songs are chosen for a prediction task, DT and NB perform consistently better than ANN and a random model. From outcomes presented in Figures 2-a and 2-b, we comfortably conclude that terms and their frequencies have predictive power to identify songs in each generation.

Visualizing Word Clouds, Terms Frequencies and Patterns

Next we like to identify the most frequent terms in each year, decade (or period), or 50 years. We also like to identify representative terms unique in the lyrics of iGeneration and other generations. To this end, we construct word clouds with terms extracted from the lyrics of 100 songs in each year. After sequentially browsing all the word clouds, we find that songs of succeeding generations use more words than those in prior generations (refer to Figures 3-a and 3-b). Note that we present only two word clouds from 1965 and 2015 due to space limitation. Note also that the size of each word in word cloud is proportional to the frequency of each word in lyrics, meaning that terms in a larger font represents more frequent terms. We observe several common key terms (e.g., “love” and “baby”) in each year repeatedly. At the same time, we observe a set of distinctive patterns over years, which will be discussed later.
Peeking Into Minds of iGeneration via Lyrics

Twenty-fifth Americas Conference on Information Systems, Cancun, 2019

To quantify our visual findings, we measure the length of lyrics (i.e., the number of letters excluding spaces) and the number of terms in the lyrics of all songs in each decade and present them with 95% confidence interval in Figure 4. It is very evident that these two metrics in the last two decades of iGeneration (Y2000s and Y2010s) are statistically significantly different from those in the other two generations.

Next, we present 50 most frequent terms in lyrics over 50 years in Figure 5. The most frequent term turns out to be “love” with 13,908 occurrences in lyrics in the past, as expected before analysis. The next most frequent terms are “im” (10,986 times), “up” (8,824 times), “baby” (6,013 times), and “youre” (5,303 times). Overall, the frequencies of most popular terms follow the Pareto distribution, roughly showing that 20% of the most popular terms occupies about 80% of the entire frequencies. For notational convenience, we present the frequency pattern of a term $t$ as follows: $t$ (frequency in 1965-highest frequency in any year-last
frequency in 2015). For example, the frequency pattern of “love” term can be denoted as “love” (249-431-301), indicating that it occurs 249 times in 1965, 431 times in peak (in 1979), and 301 times in 2015. Overall, the “love” term occurs about 300 times in each year over the past 50 years with a sharply increasing pattern in Generation X, a fluctuating pattern in that of Generation Y, and a decreasing pattern in Generation Y as shown in Figure 6.

![Figure 5: Most Frequent Terms in Lyrics](image)

The second most frequent term, “im” follows a (43-514-370) pattern: its frequencies exponentially grows during Generations X and Y starting with 43 occurrences in 1965 and peaking with 514 occurrences in 2011, but finally fluctuating around 370 occurrences on average in iGeneration. Similar patterns with different magnitudes are also observed for other popular terms such as “ill” (73-151-103), “youre” (35-200-119), and “em” (12-128-34) (refer to Figure 6). One important observation we made is that the frequencies of these four self-focus terms (e.g. “Im” and “Iill” that are related to first or second person singular pronouns) are in a steadily increasing trend and terms describing companionship and social contact (e.g. “we” that are related to first person plural nouns) are less frequent in lyrics during iGeneration, confirming the observations made in a prior study (DeWall et al. 2011). However, it warrants follow-up studies to support the finding that popular song lyrics in iGeneration indeed reflect more terms of individualistic traits from societal, psychological, political and economic perspectives.

![Figure 6: Frequency Trends of Self-focus and Gender Pronouns](image)
We also presented the trends of gender related pronouns such as “boy” (21-147-22), “girl” (57-240-86), and “baby” (97-210-127) in Figure 6. Between “boy” and “girl” terms, “girl” term has been used more often, although this finding does not necessarily provide sufficient information to infer about the issues of gender inequality and the sexualisation of female singers (Whiteley 2013; Wapnick et al. 1997). This is mainly because, most times, it is socially acceptable to call women of any age “girls,” but it’s not as common to refer to an adult man as a “boy.” Most of all, a more neutral term, “baby”, has been more heavily used in lyrics than “boy” and “girl”, and all these terms show steadily increasing trends over years and reach the peak in early 2000s.

![Figure 7: Frequency Trends of Positive and Negative Terms](image)

One of the most interesting observations was made from the frequency pattern of “up” (43-514-376), a frequency pattern curve with the steepest up-slope over 50 years. In particular, it becomes the most frequent term in the period of iGeneration. This finding could be partially attributed to the fact that many rap and hip-hop sons often include reactive tokens like “up” as in “put your hands up” in their lyrics. Another positive word, “way” (34-211-61, not shown), shows almost identical pattern as “up.” Two other words in the positive category, “wish” (10-61-10) and “dream” (11-70-4), also shows a sharply increasing frequency pattern until 1990 followed by decreasing patterns until recently. Overall, the frequency trends of these terms that represent “hopeful” mind for the future indicate that while people strongly reflected their “dreams” or “wishes” to move “way up” in their social and/or financial standings in the period of Generations X and Y. However, starting from the early period of iGeneration, they did not show such hopeful mind sets any more as strongly as they did in prior generations.

Instead, we observed that negation terms such as “aint” (43-216-107), “cant” (53-254-100), “wont” (13-110-60), “na” (18-249-139), and “nothing” (9-55-44) show steadily increasing frequency trends with peaks around early 2000s, the beginning of iGeneration as shown in Figure 7. We subjectively attribute this finding to the fact that serious challenges such as political (e.g., 9-11 event, 2001), financial (e.g., the Great Depression, 2008), and societal (e.g., ever increasing income gap between middle and upper classes) difficulties in the period of iGeneration may discourage many people and boost pessimistic minds. Along these lines of trend changes, we also observe that sex and curse terms like “bitch,” (2-87-87) “fuck,” (1-89-89) “ass,” (1-55-39), and “sex” (1-29-7) appear much more in iGeneration period than in two prior generation periods. While these terms are not listed within top 50 most frequent terms, they appear in lyrics starting around or after 1975 and their frequencies increase by 3800% to 5400% by 2015. This finding is consistent with the finding that rap and hip-hop music popular in iGeneration inherently reflects issues on violence, sex, drugs/ alcohol, racial politics, and the modern man’s struggle (Rose 1994).

**Limitations and Future Work**

In this study, we extract and observe several patterns from the lyrics of more than 5000 songs. In particular, we try to identify textual patterns observed in lyrics of songs popular during iGeneration against songs...
popular during two other generations, Generation X and Y. Our analysis confirms people’s belief that the “love” term is the most popular term across all generations although its frequency displays a decreasing pattern during Generation Y and iGeneration. In addition, we also observe steadily increasing frequency of terms related to singular first person pronouns but decreasing frequency pattern of terms describing companionship and social contacts in the lyrics of popular songs in iGeneration. We also observe lower frequency of hopeful terms (e.g., “up”, “dream”, and “way”), higher frequency of negation terms (e.g., “ain’t”, “cant”, “wont”, “na”, and “nothing”) and curse terms (e.g., “bitch,” “fuck,” and “ass,”) in the lyrics of popular songs in iGeneration.

For future work, it is necessary to take a follow-up research to re-confirm or re-test findings presented in this study. For example, we like to analyze dominant emotional sentiment of songs in each generation based on the frequencies of positive and negative terms in the lyrics of songs. In particular, we intend to investigate psychological, societal, political, and economical characteristics of iGeneration to explain the distinct patterns observed from this study.

REFERENCES


Peeking Into Minds of iGeneration via Lyrics