The Thesis Committee for Arnold Dongwoo Chung
Certifies that this is the approved version of the following thesis:

Brand Personality Research on Twitter

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

____
Gary Wilcox

____
Patricia Stout

Brand Personality Research on Twitter

by

Arnold Dongwoo Chung, B.S.; M.Int'l.Studies

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Arts

The University of Texas at Austin

May 2015

Dedication

I really thank to my parents. Without their love and support, I couldn't finish this research. Also, I really appreciate Dr. Wilcox and Dr. Stout's encouragement and support.

Abstract

Brand Personality Research on Twitter

Arnold Dongwoo Chung, M.A.

The University of Texas at Austin, 2015

Supervisor: Gary Wilcox

Social media has become a new channel for both brands and social media users. On social media channels, not only does a brand provide messages to their followers, but also social media users consume, contribute, and create brand related messages. In these social media messages, the brand personality that consumers actually do think and feel is included. In previous brand personality research, surveys have been the primary research methodology, however in this study, text mining in social media was utilized to examine brand personality. More specifically, Twitter messages, including the keywords Apple, Samsung, iPhone, and Galaxy, were collected and examined.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

In last two decades, the Internet has emerged as a technology that has changed people's daily lives. The Internet has connected people, closing the limitation of space, and more recently people on the Internet have closely connected with each other using a platform called social media. Social media allows people to engage in mass communication on their own network and allows brands and companies to reach their customers more effectively and frequently.

Each social media platform has different characteristics and strengths. For example, Twitter is good to share short messages due to its 140-character limit; Facebook suggests the people you may know; YouTube is specialized to share video clips; and Instagram provides a place to share pictures. Each platform can be an excellent channel for sharing information because of their special characteristics and strengths.

A widely discussed benefit and strength of social media would be e-word of mouth. Social media allows users to pass information they have and seek useful and interesting information. The potential spread of e-word of mouth through social media would be beyond our imagination. In 2014, Ellen DeGeneres's tweets with a picture of celebrities at the Oscars was retweeted 3.3 million times and 8.1 million people saw this picture via retweets, and the estimated total views was 32.8 million (Fleischman, 2014). From this perspective, social media such as Twitter has infinite potential power of word of mouth and reach of messages.

Since the introduction of social media platforms, their popularity has soared every year. For example, Twitter, which launched its service in 2006, had 288 million monthly active users by the end of September 2014. More people are using social media. According to Pew Research Center, about 74% of U.S. adults online reported that they are using social media and 46% of social media users post their lives and experiences on it (Duggan & Smith, 2013). On social media, users share their daily life, experience, and knowledge. In response to the wide use of social media, brands are advancing to social media webpages. According to research by the University of Massachusetts at Dartmouth, 83% of Fortune 500 companies are using Twitter and 80% of Fortune 500 companies are using Facebook (Barnes & Lescault, 2014). Those companies actively interact with their customers and engage with their followers with product and brand information, customer care, and relevant brand activities that brand followers may wish to involve in. On social media, brands may wish to find a differentiated way of marketing through engaging with followers.

As brands advance to social media, users begin to follow the brand's social media account and obtain information about the brand. On the other hand, brands on social media wish to spur engagements from their followers by providing diverse activities and disseminate their product information and brand promotions. The brand engagement may generate more followers in social media and may strengthen the relationship between the brand and follower. Furthermore, these followers of a brand in social media may provide another chance for messages to spread to followers' networks via the sharing functions available on social media. When followers of a brand become more engaged with a brand, they become evangelists of the brand and spread e-word of mouth. In the positioning of a brand, social media has become an important message channel to deliver not only products and services, but also the brand itself. Therefore, more companies are actively engaging in social media. However, there are some companies that do not use any social media at all, such as Apple.

Apple is one of the most valuable brands on earth, but it is not utilizing social media (Interbrand, 2014; Langer, 2014). Apple has some accounts in Twitter, @iTunes,

@iTunesMusic, and @iTunesTrailers, but they do not have their official brand account to represent their brand. Apple's CEO Tim Cook is using Twitter, but it is hard to say he is truly representing the Apple brand. On the other hand, Apple's biggest rival, Samsung, has 10 million followers on Twitter and actively interacts with consumers on social media. These two rival companies have competed for a long time in every aspect, but why are they taking different communication strategies on social media? Why does Apple not utilize social media? In Apple's case, it is difficult to answer the question without any information about its internal social media communication strategy; Apple must have strong supporting reasons for not using social media. On the other hand, in Samsung's case, it considered the benefits of using social media as a communication channel and place of engagement with their consumers; therefore, it is actively using social media.

Based on these two brand presences and activities in social media, I had the following questions about the two brands' social media strategies: Why doesn't Apple use social media and why does Samsung actively use it? Is it because Apple is the most valuable brand and Samsung is second in the smartphone industry? Generally speaking, what would happen if one of two competing brands is actively using social media and the other is not? What is the role of social media in general among large brands such as Apple and Samsung? What would brands expect as an outcome from marketing efforts in social media? In addition to the intended effects such as changes in purchase behavior, intention to purchase, or changes in attitude toward the brand (Thorson & Rodgers, 2012), is there a unique intended effect from social media?

On the contrary, how would consumers perceive the brand messages and information in social media in addition to traditional message exposure? Moreover, when consumers are exposed to other social media users' generated content about a brand in addition to a brand's messages; what would be the result of exposure to brand messages and user-generated content?

As a result of social media activities, two effects are expected: the change in brand image and brand personality in a consumer's mindset and users' engagement. Social media user's engagement with a brand is expressed by writing a message about a brand, by sharing the message through the "Retweet" or "Like" function, or by replies to the original source message. Thus, consumer's engagement in social media may generate diverse brand relevant topics; then, how would they describe a brand?

PURPOSE OF THE STUDY

In this study, how a brand is described on a social media channel will be tested and studied. More specifically in this study, how Apple and Samsung brands were described by consumers will be studied using a technique called text mining. According to Marti Hearst (2003), text mining is "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources." In previous brand studies, very few articles utilized text mining based on actual social media data. That is because social media is a very new media channel and text mining from social media data has not been fully adopted and utilized in the principles of communication and social science area. Collecting social media data such as Twitter text and utilizing text mining to find out new insights will be adopted in the future.

There are several reasons to choose Apple and Samsung to track on Twitter. First, these brands are the most valuable in the world; Apple is the most valuable global brand and Samsung ranked 7th among the top 100 global brands (Interbrand, 2014). Second, they have led the smartphone industry and have introduced state-of-the-art technology every year; they have strong and loyal fans in the smartphone market and their fans express their engagement on social media. Moreover, smartphone and Twitter usage by age group may have a common ground: the younger generation is more likely to use both a smartphone and Twitter. A smartphone is more likely to be owned by young people and 67% of people with a smartphone are using it every day; 37% of Twitter users are 18-29 years old and 25% of users are 30-49 years old; therefore, smartphone relevant brands and products might be discussed as a topic on Twitter (Duggan, Ellison, Lampe, Lenhart, & Madden, 2015; Google, 2013; Smith, 2013).

Chapter 2. Literature review

Chapter 2 reviews a variety of literature on social media, brand personality and social media engagement that provides a theoretical foundation for this research. The definition of social media and its impact and effect as a media channel are discussed. In addition to that, previous researches about brand personality theory and brand image are further discussed. Then, brand image and consumer's engagement on social media will be discussed with social media engagement.

2.1 Social media and e-word of mouth

People already know what social media is and probably may think of social media services such as Facebook, Twitter, or Instagram. There are several definitions for social media. boyd and Ellison (2007) defined social media as a "web based service that allows individuals to (1) construct a public or semi public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, (3) view and traverse their list of connections and those made by others within the system, and (4) create and share content." Similarly, Kaplan and Haelein (2010) defined social media as, "a group of Internet-based applications that build on the ideological and technological foundation of web 2.0, and that allow the creation and exchange of User Generated Content". User Generated Content is the content "published either on a publicly accessible website or on a social networking site accessible to a selected group of people"; it "shows a certain amount of a creative effort"; it is "created outside of professional routines and practices" (Vickery & Wunsch-Vincent, 2007). Thus User Generated Content is summarized as publicly available content created by social media users. Based on definitions of social media and User Generated Content, two important characteristics of social media can be highlighted: creation and exchange (sharing) of content by its users (Muntinga, Moorman, & Smit, 2011).

On social media, people can create and distribute content. Moreover, the creation of content has become simple and what a social media user can do is write text, take a picture, or shoot a video, and then a user posts it on social media (Thorson & Rodgers, 2012). Once the content is posted on social media, it is shared and communicated through personal connections that a user has made (boyd & Ellison, 2007). Also, a social media user shares his/her experience with a brand and shares the information. Jansen and his colleagues found that about 20% of tweets mentioning a brand carried sentiments about a brand/service/product (Jansen, Zhang, Sobel, & Chowdury, 2009). According to the data they collected, about 60.6% of tweets that mentioned a brand were positive sentiment tweets and 22.2% of them were negative. Besides this, Jansen et al. (2009) also found that the remaining 80% of brand relevant tweets were "seeking information, asking questions, and answering questions about a brand"; therefore, social media functions as a source of information about a brand for its users. Moreover, consumers express their feelings and share their information about a brand; inevitably some of the shared messages may include consumer's perceptions about a brand. These sharing the perception or feeling or experience about a brand may cause e-word of mouth about a brand.

E-word of mouth can be defined as interpersonal communication about products and services within the online space which may affect consumer's attitude and behaviors like word of mouth (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004; Richins, 1984). Unlike the traditional word of mouth that is spoken, person-to-person, interpersonal communication (Arndt, 1967; Richins, 1984), e-word of mouth on social media does not need to be like traditional word of mouth: e-word of mouth spreads to the network of people effectively and immediately (Hennig-Thurau et al., 2004). Moreover, in e-word of mouth studies, e-word of mouth has been found to include not only the word of mouth characteristics such as opinion leadership and opinion seeking, but also opinion passing (Chu & Kim, 2011; Feick, Price, & Higie, 1986; Gilly, Graham, Wolfinbarger, & Yale, 1998; Sun, Youn, Wu, & Kuntaraporn, 2006). In word of mouth, opinion leadership is "sharing information and influencing other's attitudes and behaviors" and opinion seeking is "looking for information or advice from others when making an informed decision or taking action". But e-word of mouth has an additional attribute, opinion seeking; opinion passing is "passing-along or forwarding of messages" (Sun et al., 2006).

Because of opinion passing, e-word of mouth is different from word of mouth. In word of mouth, opinion is spread through person-to-person communication, but opinion passing on social media can be made without it. More specifically, on social media, either the information disseminated from the network or the information they sought is further passed along to user's networks by sharing functions such as the "Like" or "Retweet" function provided from the social media platform (Cha, Haddadi, Benevenuto, & Gummadi, 2010; Jansen et al., 2009; Kwak, Lee, Park, & Moon, 2010). Furthermore, social media messages and information shared by family, friends, colleagues, and other consumers are considered as trustworthy and more credible information sources than other information sources (Chu & Kim, 2011; Foux, 2006). These shared messages from personal contacts might be further spread to other social media users and might eventually cause e-word of mouth (Lipsman, Mudd, Rich, & Bruich, 2012). Thus, it can be argued that social media users are naturally involved in e-word of mouth because they post messages, seek information, and pass along the message and information to their network.

Thus, social media is a communication channel for creating and sharing the information. In this communication channel, its users actively share information about a brand; these sharing activities may trigger e-word of mouth and promote further spread of information between users.

2.2 Modern communication, brand, and social media

According to modern communication theory, advertising works under the framework of "Who says What in Which Channel to Whom with What effects" (Lasswell, 1948). In other words, there are five players in message framework-message sources (who), a message (what), channel (which channel), recipient (whom) and what effect (intended effect) and messages from sources were carefully designed for recipients and the intended effect. In advertising and marketing studies, a brand used to function as the message source of communication and a consumer was a message receiver. Through traditional media channels including newspapers, magazines, TV, radio, or the printed ad, a brand (source) has tried to deliver brand image and personality, and offer value to consumers (recipients) (Nandan, 2005; Plummer, 1984). In this one-way communication, consumers generally consume the information and message from brand and may understand the message and create their own brand personality and image based on the brand messages.

However, social media changed the paradigm of one-way communication (Meraz, 2009). Since people can easily create and share content on social media, this traditional advertising and marketing message framework does not work on social media. The message source is no longer limited to the brand, advertiser, and marketers, but each individual social media user becomes another source of information about a brand (Jansen et al., 2009; Meraz, 2009; Muntinga et al., 2011). These messages about a brand are "publicly available" for other users to "consume, comment on, or further modify" (Gangadharbatla, 2012; Vollmer & Precourt, 2008) and are out of control from the perspective of a brand (Mangold & Faulds, 2009). Therefore, Mangold and Faulds (2009) argued a brand and brand managers "should recognize the power and critical nature of the discussions being carried on by consumers using social media." Furthermore, they argued that a brand must "shape the discussions" on social media by providing tools to engage customers, by providing a networking platform, by providing information, etc. (Mangold & Faulds, 2009). However, understanding reasons that social media user's motivations to create and share the messages about a brand are necessary.

2.3 Use and Gratification theory, consumer online brand related activity, and social media engagement

In mass communication studies, use and gratification theory examines how and why people use and seek media based on the assumption, "people actively seek and use media" (Ruggiero, 2000). Since use and gratification theory argues about certain motivation that drives consumption of certain media, for example mobile TV or the Internet, and its consequent satisfaction from media consumption (Ko, Cho, & Roberts, 2005; Kyun Choi, Kim, & McMillan, 2009), people's use of social media as a communication channel could be understood within the use and gratification theory. According to the use and gratification theory, there are four categories of motivations to use a certain media: entertainment, integration and social interaction, personal identity, and remuneration and empowerment (McQuail, 2010). Based on these four categories of motivations, Muntinga et al. (2011) suggested the Consumer Online Brand Related Activity Model (COBRA).

According to the COBRA model, consuming, contributing, and creating are three activity types of brand-related content on social media and social media users consume, contribute, and create brand-related content based on their involvement level (Muntinga et al., 2011). Based on McQuail's classification (McQuail, 2010), Muntinga et al. (2011) suggested COBRA motivations were entertainment, integration and social interaction, personal identity, information motivation; but each type of COBRA has different mixture of motivations. For example, when a social media user decides to contribute to COBRA, social media user's motivations-personal identity, integration and social interaction and entertainment-spur contribution activity on social media such as a reply, retweet or like on social media. On the other hand, personal identity, integration and social interaction, empowerment, and entertainment played an important role for creating COBRA on social media.

In contrast to mass communication's attempt to understand the media user's motivation, effects and effectiveness of social media usage and brand marketing mix on social media have been studied in the field of advertising (Thorson & Rodgers, 2012).: e-word of mouth (Chu & Kim, 2011); purchase decision making (Powers, Advincula, Austin, Graiko, & Snyder, 2012); further reaches of brand messages by fans' engagement (Lipsman et al., 2012); purchase intention (Van-Tien Dao, Nhat Hanh Le, Ming-Sung Cheng, & Chao Chen, 2014; Wang, Yu, & Wei, 2012). In addition to the intended effects, advertisers examine the effectiveness of advertising in social media by measuring how many times the advertising message was shared by the audience, how many reviews were added to the messages, or how many times external sources in the message were clicked. All of the activities undertaken by social media users after exposure to the message on social media are called engagement (SAS Institute, 2010; Solis, 2010).

An SAS Institute white paper (2010) described engagement as "…some evidence…at least somewhat interested in the brand" and it argued that advertising needs more engagement rather than frequency and reach. On the other hand, Brodie, Hollebeek, Juric, and Ilic (2011) argued that "consumer engagement in a virtual brand community involves specific interactive experience between consumers and the brand, and or other members of the community." Based on these definition, it can be argued that social media user's engagement is evidence that social media user is interacting with a brand or other users. Also, it can be argued that social media user's engagement with a brand is relevant or equivalent to COBRA's contributing or creating activities since social media users can interact with a brand or other users through contributing and creating of contents. These contributing or creating activities left in the form of a retweet, reply, or message on social media.

2.4 BRAND IMAGE AND BRAND PERSONALITY

Since the failure of consumer personality segmentation, marketers and advertisers accepted brand personality concept and utilized it into their marketing mix (Plummer, 1984). Advertising and marketing field practitioners used brand personality "to differentiate their products and brands", "to create emotional aspects of a brand", and "to argue the personal meaning of a brand to the consumer" (J. Aaker & Fournier, 1995; Crask & Laskey, 1990; Lannon & Cooper, 1983; Levy, 1959).

According to Plummer (1984), there are two-dimensions in brand images and two-dimensions in brand personality. First, Plummer (1984) argued that brand images are the images that "a brand presents itself to the world" and the images that "world interpret the brand through many different filters" including experience, perception, misconceptions, the value systems, and all the noise. In addition to that, Park, Jaworski, and Maclnnis (1986) argued that brand image is "not simply a perceptual phenomenon affected by the firm's communication activities alone," but it is "the understanding consumers derive from the total set of brand-related activities engaged in by the firm."

Plummer argued a brand image could be described in three ways, "physical attributes", "functional attributes", and "characterizational aspects" that are brand personality.

In contrast to brand image, Plummer (1984) argued that brand personality has two dimensions: "input as what we (advertisers and marketers) want consumers to think and feel and out-take, what consumers actually do think and feel." Comparing to brand image, brand personality provides and delivers more solid characteristics of brands that a brand wish to present to a consumer. According to Plummer (1984), advertising and marketing practitioners utilize the brand personality in two ways; they create brand personality statement to create the advertising and brand personality profiles to measure and evaluate perceptual reality from the consumer perception. In other words, brand personality profiles are the measurement of intended effects from brand personality input such as advertising.

Contrary to Plummer (1984), Aaker (1997) defined brand personality as "the set of human characteristics associated with a brand." Based on a questionnaire survey asking which adjective describes the brand well by providing the list of adjectives, Aaker proposed that there are five brand personality dimensions: sincerity, excitement, competence, sophistication, and ruggedness. Since Aaker's manuscript about brand personality, several scholars have tried to argue against and develop a new brand personality dimension by substituting one or two dimensions of brand personality (D. A. Aaker & Joachimsthaler, 2000; J. Aaker & Fournier, 1995; Jennifer Lynn Aaker, Benet-Martinez, & Garolera, 2001; Bosnjak, Bochmann, & Hufschmidt, 2007; Caprara, Barbaranelli, & Guido, 2001; Geuens, Weijters, & De Wulf, 2009).

A common ground of these brand personality researches is that researchers provided the list of adjectives to survey participants and ask them to match or describe the brand based on the list. Biel (1992) argued, "A good starting point is to describe the image of a brand as that cluster of attributes and association that consumers connect to the brand name", and Aaker (1997) and other scholars followed Biel's methods and tested it with brand personality. Also, brand personality theory only tested how survey participants think and feel about a brand; previous brand personality studies should rely on survey and participants' responses. However, through utilizing social media, researchers can observe and collect the data that how people think and express about a brand.

2.5 SUMMARY OF LITERATURE REVIEW AND HYPOTHESIS

The introduction of social media changed the paradigm of one-way communication; now people easily create and share the information on social media. These people online may cause e-word of mouth and social media users actively consume, contribute and create brand related content. Before and after exposing to brand messages and user generated contents about a brand, consumers may have a brand image and brand personality. Through engagement (or COBRA) on social media, each consumer's brand personality could be found.

Based on the research questions and literature review, following is my hypotheses to be tested using Twitter tweet collection with keyword Apple, Samsung, iPhone, and Galaxy:

H1. Consumer Online Brand Related Activity will be detected on social media.

H2. Some of social media user's messages (COBRA) describe brand personality.

H3. Some adjectives will be used more than others in brand relevant tweets on social media, which is brand personality.

H3-1. In the case of Apple and Samsung, and the iPhone and Galaxy smartphones, adjective terms used by social media users will be different for each keyword.

H4. Aaker's brand personality adjectives will be found in brand relevant tweets.

H4-1. In the case of Apple and Samsung, and iPhone and Galaxy, Aaker's adjective terms to describe the brand personality will be included in the descriptive terms and they will be different for each keyword.

Chapter 3. Methodology

The chapter 3 reviews a methodology implemented in this research. More specifically, how Twitter data was collected and how Twitter data was cleaned with filters will be described and articulated.

3.1 INTRODUCTION

In order to test the research questions and understand the consumer on Twitter, an analysis of tweet content should be employed. To conduct the analysis of tweet content, the author should have reasonable amount of tweets relevant to the research topic over the period of data collection. Perhaps only a few hundred or a few thousands of tweets might be collected through the search function provided on Twitter, but it returns only few tweets generated and the Twitter algorithm prevents to capture currently generated tweets from other users. Constant and well-representative data should be collected on Twitter. Also, collecting the right data is necessary. 500 million tweets are generated each day and we may not know how many of them are relevant to the research topic.

For the social media content research, data collection from social media is the starting point of research. More specifically, data should be stored in the physical hard disks for future data analysis; data should be in a readable format and should be loaded on a computer for further analysis; data should be wrangled and cleaned by filter and purpose; and final data should be visualized as the result of data analysis (Figure 3.1). Computer programs should run all of these processes with great care.

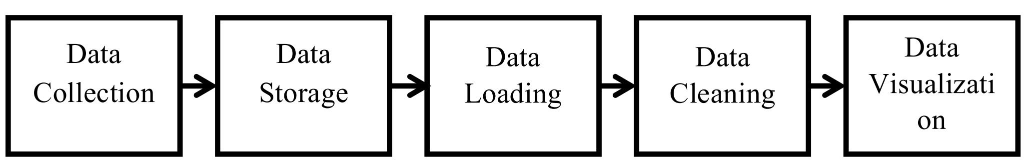

Figure 3.1 Social media data management process

In the field of social science, there are not many studies related with Twitter or other social media such as Facebook or Instagram, and not many studies actually used real social media data for their research analysis (Jansen et al., 2009; Tumasjan, Sprenger, Sandner, & Welpe, 2010). There are many programming languages such as PHP, JAVA, Perl, Ruby, and Python that allow collecting Twitter data via Twitter APIs. Among them, the author chose Python for these reasons: it is free; it is easy to use and learn; the community (Stack Overflow) for information; free libraries available online such as Natural Language Toolkit (NLTK) and Tweepy for research purposes. For this research, Twitter data collection was conducted based on four keywords through computer program coding and collected for 115 days from November 10, 2014 to March 5, 2015.

3.2 DATA COLLECTION

For data collection, Python 2.7.9 was chosen and utilized. Python library Tweepy and streaming APIs were utilized for Python data collection scripts. To discuss Twitter data collection further, understanding Twitter APIs, rate limits, and OAuth of Twitter should be necessary. API is an abbreviation for Application Programming Interface and it is “a hook for colleagues, partners, or third party developers to access data and service to build application” (Jacobson, Woods, & Brail, 2011). In other words, it’s a set of functions for communicating with the Twitter HTTP server. Rate limits are the restrictions of using Twitter APIs set by Twitter to avoid abusive use of Twitter APIs and the Twitter server. OAuth is the way Twitter provides data to the third party. More elaboration on the APIs, rate limits and OAuth will follow.

3.2.1 REST APIs

There are three public APIs documented on Twitter' developer's sites: REST APIs, Streaming APIs, and Ads APIs. But, functions of REST APIs and Streaming APIs should be discussed. According to the Twitter's developer webpage, REST API is described as:

> "The REST APIs provides programmatic access to read and write Twitter data. Author a new Tweet, read author profile and follower data, and more. The REST API identifies Twitter applications and users using OAuth; responses are available in JSON." (Twitter Inc., 2015i)

As Twitter described, REST APIs allow a user/developer to read and write Twitter data. A Twitter developer requests the Twitter HTTP server to do functions with REST APIs, then Twitter HTTP server returns the values of requests based on user API requests (Figure 3.2). However, REST APIs have certain limitations to use such as rate limits which allows only 15 requests per 15 minutes. For example, you make a request with the API, GET status/user_timeline, and try to read tweets on CNN (@CNN) timeline. This request is counted as one request, and it will return up to 3200 tweets of a user's most recent tweet (Twitter Inc., 2015r). Each user can make 15 requests per 15 minutes; if a user used all 15 requests, he/she needs to wait 15 minutes for a reset of the remaining API counts (Twitter Inc., 2015a). Using REST APIs may be useful when a researcher knows only when small numbers of past tweets are needed. REST APIs only returns up to the most recent 3200 tweets that are stored on the Twitter HTTP server. Moreover, since June 2013, a Twitter user cannot collect to go beyond certain numbers of paginations through the REST API (Twitter Inc., 2015r). To solve the rate limit and pagination limit, we need to utilize the streaming APIs.

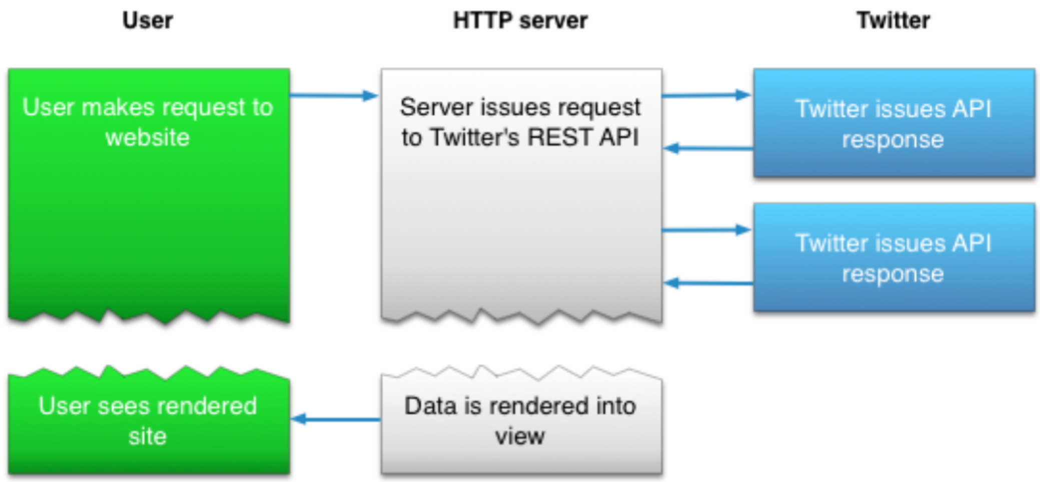

Figure 3.2 REST APIs (Twitter Inc., 2015n)

3.2.2 Streaming APIs

> "The Streaming APIs continuously deliver new responses to REST API queries over a long-lived HTTP connection. Receive updates on the latest Tweets matching a search query, stay in sync with user profile updates, and more. If your application is rate-limited for over-polling the REST APIs the Streaming APIs may be a good solution for your needs." (Twitter Inc., 2015c)

According to the Twitter description, streaming APIs "deliver new responses to REST API queries over long-lived HTTP." In other words, unlike REST APIs which a developer makes a request to read and see the data on Twitter server, streaming APIs allow a developer to request Twitter contents over a long period of time after they make a connection to Twitter HTTP server (Figure 3.3). Until the connection is closed or an error flag rises, it will push the tweets based on the Streaming APIs requests that users set. Streaming APIs requests can be a "track" of keywords, or "follow" of users, or "location" of users, or "language" of tweets. (Twitter Inc., 2015k)

In this research "track" of keyword (or phrase) is used because other streaming requests will not provide the specific data. Streaming API request "follow" will return a specific user's Twitter activities such as tweet post by user, retweets by users, replies to tweets by users, manual replies by users. Streaming API request "language" will capture the specific language such as English, Japanese or Chinese; it is practically impossible to catch all English tweets due to limitations of streaming APIs. The location of users may provide tweets generated in the United States, but users can opt out of their location; there are chances we may lose tweets by opt-out users. Therefore, tracking certain keywords would guarantee how a brand is described.

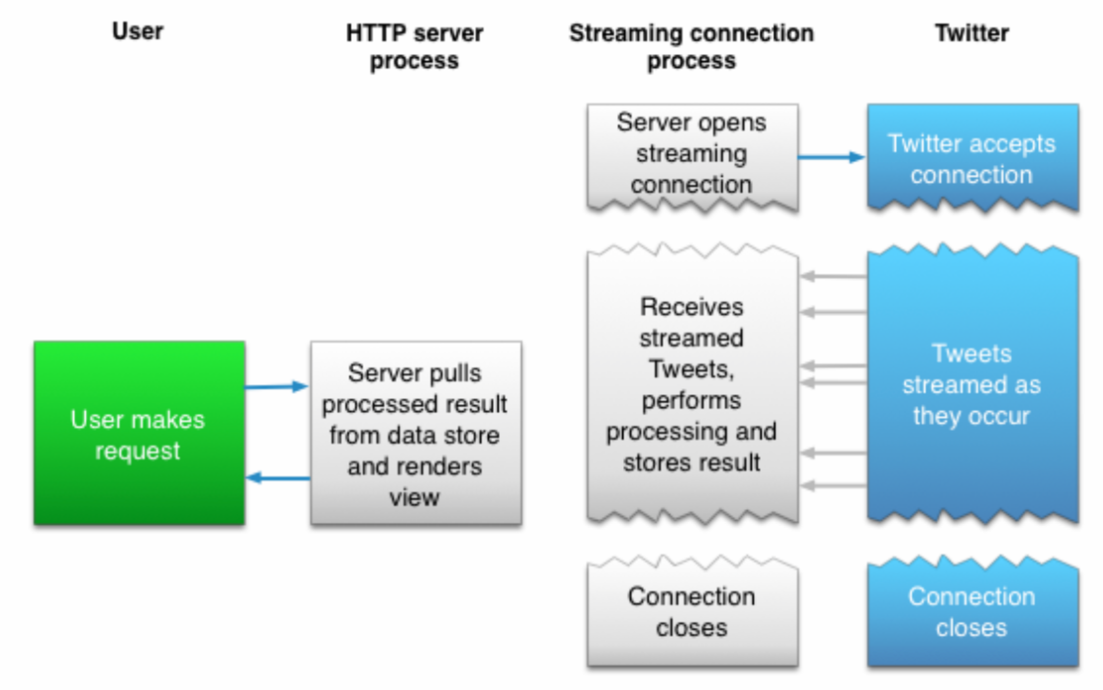

Figure 3.3 Streaming APIs (Twitter Inc., 2015n)

Even though streaming APIs provide unlimited numbers of tweets for data collection, there are certain weaknesses: track method, live-streaming, and returned value. First, the weakness of the streaming API is a limitation in the track method that researchers have to use. To obtain certain keyword relevant data, the "track" function of streaming APIs should be used. Track is defined as "A comma-separated list of phrases which will be used to determine what Tweets will be delivered on the stream" (Twitter Inc., 2015k). The search term is set as *'search term'* or *"search term"* in either quotation marks or double quotation marks in the list of search terms so that streaming APIs will filter the "search term".

One important thing to remark about in using track Streaming APIs is that streaming APIs push everything that matches the condition. In computer data structure, the phrase inside the quotation mark is called a string. The Twitter HTTP server pushes not only all strings in the quotation mark, but also strings which includes the search term string. More specifically, if a string in the track list is a substring of a bigger string, the bigger string will be caught and will be returned. For example, the keyword "search term" in the track list will return, "search terms" because the keyword "search terms" is a bigger string and includes "search term" as a substring. Moreover the space between terms in a phrase causes some inaccuracy because the Twitter HTTP server considers space between terms as "and". For example "search terms" returns the all the sentence that meets the condition search and term such as "search relevant term" or "relevant search term." Therefore when streaming APIs were used, phrases in the track list should be a very specific expression for accuracy of returns (Twitter Inc., 2015k).

Second, the data returned from the Twitter HTTP server may not be the whole currently generated tweets in Twitter. Twitter traffic volume is changing every second, and volume may soar at the moment of an important event. If the phrase of the tracking list is generating Twitter traffic volume bigger than 1% of firehose capacity, then pushed data from the Twitter HTTP server will be a sample of currently generating data and some tweets will be missed (Singletary, 2012).

Third, since it only pushes currently generated tweets, there is no way to go back and collect with Streaming APIs. Collecting tweets on an individual computer may have a moment of electricity outrage or an unstable Internet connection, which may cause the disconnection to streaming APIs. Therefore, a stable connection of streaming API is required for data collection.

3.2.3 OAuth

Some webpages require users to register membership and ask for a user ID and password to access their services. As the number of Internet service providers grows, there is a need to access the data of an Internet service from a third party to provide another service based on the Internet service. For example, recently, some webpages can log a user in with their Twitter account, Google account, Facebook account or other major Internet services. When a user of a webpage clicks to log on the Internet service with Facebook, the webpage requests Facebook to access the user data and collect the data to create the credential for its webpage. However, in the position of Facebook, Facebook does not wish to allow the third party (outside Internet service provider) to access their whole data set and collect the whole user profile, but wishes to allow only part of their data to use. Therefore, an Internet service provider such as Facebook, Twitter, Google, or others needs to have another way of allowing a third party user to access part of their data, which is OAuth.

Twitter OAuth is used to grant the authority for third party to access to part of the Twitter data set. In the above example, the webpage asks Twitter about user data access and if the condition is met, Twitter provides user information to the third party. Similarly, OAuth grants authority to access their past Twitter data, streaming data, user information, timeline status, etc. To obtain access, each researcher and developer needs to create his/her own app and agree to the terms of use at the Twitter’s developer’s site (https://apps.twitter.com). Twitter OAuth consists of four elements: Consumer Key, Consumer Secret, Access Token, and Access Secret Token. Providing these four OAuth keys, a researcher can access the Twitter data through OAuth and can collect data through APIs.

3.2.4 API Rate Limit

To use the REST APIs and Streaming APIs, Twitter users must be aware of the API Rate Limit. API Rate Limit was introduced to prevent the abuse of Twitter APIs (Twitter Inc., 2015i). If a user repeatedly abuses the APIs, Twitter blacklists the user and that IP is blocked to use Twitter. Also, due to the API Rate Limit, past tweet mining becomes unavailable because the API command to obtain tweets is limited.

3.2.5 Data collection

For data collection, Tweepy is utilized among free libraries (Roesslein, 2014; Twitter Inc., 2015p). Tweepy is a module used to extract data using OAuth and streaming APIs. Utilizing the Tweepy library, users can obtain Twitter data in the format of JSON. The data collected from Twitter should be stored on a hard disk and should wait until enough data are collected.

3.3 Data storage and JSON

Data was encoded in the JavaScript Objective Notation (JSON) format and pushed by the Twitter HTTP server (Twitter Inc., 2015g). The JSON format data consists of keys and values in a double quotation mark; each key has a corresponding value. Each JSON data can include a data construct such as a string, array, numbers, object, or Boolean ("Introducing JSON," 2015). In each tweet, it includes keys including “created_at”, “id”, “id_str”, “text”, “source”, “geo”, “reply count,” etc. and each keys has its own value. For example, the key *"created_at"* has its corresponding value "*Sat Nov 29 04:38:27 +0000 2014*."

The benefit of the JSON format data is its ease of data management. If a researcher knows what kinds of data he needs, then he can extract the value of data by calling the key in the JSON. In this research, only text of each tweet is pulled; calling JSON key 'text' of each tweet returns value of tweet, for example, "*#iPhone #App #iSpeak #and #Spell - #Entertainment #enable #javascript #video #news #see #application #ios #6.1... http:\/\/t.co\/lvlXTblviB*."

3.4 Data loading

For the data analysis, data stored on the hard disk should be loaded onto Python. Initial raw data was a mixture of keyword data; each file size was about 100 megabytes. About 90 files were created to store the data in each day. Since the overall data size of this research was too big, raw data was split and sorted by the date; first file of each date was the starting date. Each day, about 1.8 million tweets related with the topic were collected. Initial attempt for analysis was weekly analysis based on keywords. However, in the case of Apple and iPhone, the size of each keyword-sorted files were too big to load on a personal computer; in other words, they were discussed more on Twitter than other keywords. Data was split and loaded by date so that further analysis became available.

3.5 Data cleaning

After data loading, data cleaning were done. Even though the data had been successfully collected from Twitter, the biggest challenge of this research was how to clean the data. The purpose of this study is how consumers describe a specific brand (Apple and Samsung) and products (Galaxy and iPhone) on Twitter in English. Since

Apple and Samsung products are sold in almost every country, Apple and Samsung are discussed in diverse languages. These diverse languages other than English should be excluded from the data set for further analysis. In addition to language cleaning, hashtag, Twitter ID, retweet mark (RT) in front of tweet message, and external source link (http) should be handled for text analysis. Polysemy of keywords should be excluded based on the topic and context. Furthermore, noises included in keywords should be removed from the data set.

Since the data collected for this research exceed millions, an individual content check was practically impossible. A rigorous and repetitive test of Python code was conducted with sample data.

3.5.1 Language filter

A language filter was applied to the raw data. To filter out Chinese, Japanese, Spanish or many other non-English languages, the author filtered the raw data three times: returned tweet language value, computer language encoding system, and English Dictionary.

First of all, Twitter detects tweet language by itself and determines which language is used. If Twitter determined tweet text message language, it gives values to JSON key, "lang," based on ISO 639-1 language code. In the case of English, ISO 639-1 code is "en"; therefore, "lang":"en" should be included in English tweet message (Twitter Inc., 2015e). When loading the raw data of tweets, the author set the filter and made Python to collect tweets having "lang":"en" only.

Even though Twitter detects and assigns language value, it often fails to detect the right language when tweet text is a mixture of different languages. For further accuracy in language detection, the ASCII encoding method was applied. ASCII, American

Standard Code for Information Interchange, is 7-bit encoding scheme that supports 128 characters ($2^7 = 128$) including uppercase and lowercase alphabets, numbers, punctuation, control codes, and space (Brandel, 1999). If tweet text is written in languages other than English, the ASCII encoding method cannot express non-English characters and will not further process the tweet text analysis. However, when other alphabet-based language such as Spanish, French, or some other European languages are written in a tweet message, this language filter may not function well.

Finally, an English dictionary was utilized. Each word in the sentence is compared to the dictionary. If word is not in the English dictionary, that tweet will be removed from the list of text for further text analysis. Two English word list text files were downloaded from the Internet and merged (Lawler, 1999; SIL International, 2005). In addition to these two English word lists, brand name Samsung, product names iPhone and Galaxy S, other major smartphone manufacturers' brand names were added on the list. This approach may miss the tweets which are misspelled or miss the text which are written in twitterian jargons. However, this method provides realistic way for interpretation of text.

3.5.2 Hashtags, user ID, external links, and retweet in tweet

For text analysis, hashtag, user id, external link and retweet were cleaned.

Hashtag is a word or unspaced phrases that is written after number sign #. In social media, a hashtag is usually used to share their message with a public audience and to be grouped in a tagged message. For example, there was a tweet,*"#Apple edges out #Samsung in global #Smartphone sales,"* has three hashtagged words. This tweet message can be found at all of #Apple, #Samsung, and #Smartphone and further tweet discussion with other Twitter users. Interestingly, some of these hashtagged keywords are used as words in the sentence instead of plain words. For text analysis, number sign # were removed, but the hashtag words used in the sentence should remain.

Unlike a hashtag, user id and link to external sources the URL in tweet text should be removed for further analysis. Twitter users add an @ sign in front of the user id in a tweet message so that they can send their tweet to other users. In addition to user id written in tweet text, external source URL should be managed. Since social media popularity has soared, more Twitter users share information from outside of Twitter including the URL link. In addition to information shared by users, recently many webpages such as the New York Times have included a tweet function on their webpage that allows readers to send a tweet and URL address of an article. A Twitter id and external source URL is not necessary for this research purpose; it was removed.

Finally, punctuation and retweet text should be managed. Punctuation other than @ and # are included in ASCII codes and character other than upper case, lower case, and numbers are removed for further simple data analysis. In addition to punctuation, retweet mark, RT, should be handled. Retweet (RT) is sharing a tweet of other users because Twitter users think that the tweet is important or beneficial to his/her following users. When a user retweets the message, an “RT” sign is added to the retweeted message. We may know how retweeted information is important by counting retweet numbers and how information is spread by checking users who retweeted, but it is not necessary for text analysis. However, the text itself is an important source of information because a user agrees to the idea of an original tweet text. Therefore removing the RT sign in front of a retweet message and including the tweet text should be done.

3.5.3 Topic filter: Topic Modeling and Word Collocation

Although foreign language, hashtag, user id, external source URL, and retweet are replaced or removed, further data filtering and cleaning is necessary. Unless a social media study has a very solid and specific term such as measles, there are chances of polysemy and other meanings. Since the Twitter HTTP server just pushes the entire tweets that include keyword at the moment of tweet posting, these words with polysemy are also pushed to the researcher's collection.

To remove the polysemy, a topic modeling module, Gensim, was adapted for analysis (Řehůřek & Sojka, 2010). Topic modeling is a mathematical and statistical method based on Term Frequency-Inverse Document Frequency (TF-IDF), Single Vector Decomposition (SVD), and Latent Semantic Indexing (LSI) that identifies patterns and relationships among terms. If a word appears with another word repeatedly, these words may have a strong relationship and may have a similar context. For example, let's assume that we have three sentence (corpora), "children loves eating apple," "eating apple provides necessary vitamins C," and "eating apple is good for health." Two words "eating" and "apple" appears repeatedly in three sentences and we may know that these two keywords have a strong relation. Moreover, we may guess these sentences have a common topic, "eating apple." Latent Semantic Indexing focuses on that fact and transformed all the words into the vector coordinates. If a word appears with another word in the sentence repeatedly, it will have similar direction on the vector space and may indicate that two words may have a similar topic. With topic modeling based on TF-IDF and LSI, the overall topics of tweet data could be understood.

In addition to TF-IDF and LSI, word collocation from the Natural Language Toolkit was utilized. Word collocation is the algorithms finding the terms, which appear together. Word collocation allows the researcher to know what terms are used together under the topic and corpora. Terms returned from word collocation have great implications because further clear meaning of words can be found. Furthermore if word collection is implied to filter the irrelevant topics, “apple cider” has a clearer meaning than “cider”. Combining topic modeling and word collocation, more accurate text filtering becomes possible.

3.5.3.a) Apple

In the case of Apple, Python returned the topics including fruit, New York City traveling, women’s sexy hip or other irrelevant topics. In case of the topic of the fruit apple, “green apple”, “cinnamon apple”, “granny smith apple”, etc. were found and food cooked with fruit apple, “apple crumble”, “apple cider”, “apple pie”, etc. were also found. These words should be detected in advance and excluded. In the case of New York City, “Big Apple” and “traveling” appeared and body image relevant terms, “apple hip”, were found. All of these irrelevant terms should be managed. After excluding the irrelevant topics other than the brand Apple, noises related with the brand Apple should be handled. Under the topic of brand Apple, irrelevant topics such as Apple iPhone application, Apple iPhone games, Apple reseller’s promotion, and other irrelevant topics should be found and sorted out further to find the tweet relevant to brand Apple. Apple application promotion messages or reseller’s messages have some patterns and they include some terms such as “win it” or “try it” or “download it” or “check out”. In addition to these irrelevant messages, there are some tweets related to the public figures and brand Apple that should be handled. For example, a pop singer Jacob Whitesides performed at the Apple store and his fans shared this experience with pictures on Twitter (Figure 3.4). This tweet is somewhat relevant to the Apple brand because he gave a performance at the Apple store, but it is not truly relevant to the Apple brand image, Apple brand personality, or Apple brand identity; therefore, such kind of tweets should be removed for further data analysis.

Figure 3.4 Different topics of Apple brand

Finally, some idioms related to Apple were found on Twitter. “The apple of eye”, “Adam’s apple”, and “Apple a day keeps the doctor away” were frequently found; they were excluded.

3.5.3.b) Galaxy

Galaxy, the smartphone brand of Samsung, has another meaning, “a large system of stars held together by mutual gravitation and isolated from similar systems by vast regions of space” (galaxy, 2015). In Twitter, Galaxy was frequently used to discuss the meaning of universe than the Samsung smartphone brand (Figure 3.5). Starting with this research, Hollywood movie, *Interstellar*, was released and people’s interests about galaxy soared. In addition to that movie, another blockbuster movie, *Guardians of the Galaxy*, prompted a huge Twitter discussion and these tweets were caught due to the keyword “galaxy.” In addition to recent release of Hollywood movies, the old classic movie *Star Wars* relevant tweets were found due to its strong network of fans on Twitter. Therefore, other galaxy relevant terms should be found and filtered.

Figure 3.5 Polysemy example of Galaxy

3.5.3.c) iPhone and Samsung

iPhone is the unique brand name manufactured and sold by Apple. However, a lot of iPhone application marketing and iPhone case sales buzz were made on Twitter. Some smartphone application developers encourage their users to send the tweet to user's social media network to promote the application downloads. Samsung, South Korean consumer Electronics Company, has a unique name and fortunately it doesn't have polysemy. However, Samsung produces all kinds of consumer electronics, there is a tremendous amount of noise that resellers of Samsung created. All of irrelevant noise should be found and filtered.

3.5.4 Robot or buzz makers filter (Abusers)

There are some Twitter accounts that are attempting to make a heavy buzz on Twitter. These accounts are assumed to be robots that are managed by some people; they generate hundred or thousand of tweets within a short period of time. The purpose of these accounts is getting attention of Twitter users by dumping the messages with famous hashtags such as #Samsung or #iPhone6 or #Apple and lead users to click the external link (Figure 3.6). Since streaming APIs push all tweets matching to the conditions, these tweets are also pushed to the data collection files and create serious data distortion due to its volume of tweets. These tweets are not retweeted but deleted by either Twitter; they cause serious data distortion due to their volumes. There are lists of robot accounts found on data and all of tweets generated from these accounts must be excluded from text analysis data. (These accounts are usually suspended or deleted by Twitter)

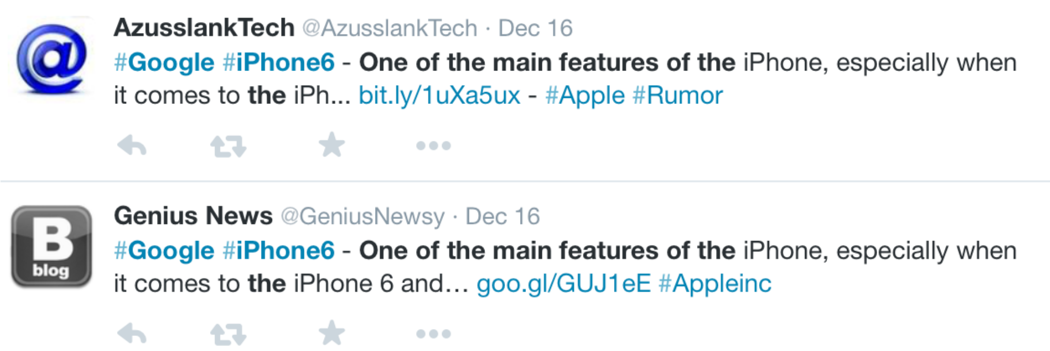

Figure 3.6 Examples of Twitter buzz makers (Abusers)

3.6 Text mining and text analysis

Social media users who were engaged with a brand may express their engagement by retweeting or sharing information on Twitter. These user-generated contents may provide some undiscovered insights about a brand by text mining. According to Hearst (2003), text mining is:

> "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation."

Text mining on social media is already adopted and utilized in diverse academia: disease surveillance, flu detection and prevention in public health (Corley, Cook, Mikler,

& Singh, 2010), public opinion mining and election prediction in political science (O'Connor, Balasubramanyan, Routledge, & Smith, 2010; Pak & Paroubek, 2010; Tumasjan et al., 2010), earthquake and forest fire detection (De Longueville, Smith, & Luraschi, 2009; Sakaki, Okazaki, & Matsuo, 2010). Also, text mining from Twitter adopted in brand research has provided some insights about what Twitter users share about a brand, as well as their sentiment towards the brand (He, Zha, & Li, 2013; Jansen et al., 2009; Mostafa, 2013).

However, the number of Twitter messages shared and the number of Twitter users has increased far beyond than when Jansen and colleagues studied their work in 2009. More users, more brands, and new features on Twitter changed the Twitter environment and it is expected that previously unknown information about a brand would be discovered through this research's text mining.

For text mining, the Python library, natural language toolkit (NLTK), was utilized and adapted (Bird, Klein, & Loper, 2009). Natural language toolkit is the Python library for natural language processing. Natural language processing is "an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things" (Chowdhury, 2003). Through the tools of natural language processing, researchers can train computer to analyze human language pattern and find the meaningful results. Since the purpose of this research is finding the adjectives that describe the brand, part of speech tagging was adopted. Based on the structure of sentence, the word class of each word is determined such as noun, verb, adjective, adverb, etc. Even though other word classes may include important information, only adjectives returned from the NLTK algorithm were collected. In addition to adjective collection, Aaker's 42 brand personality adjectives (Aaker, 1997) were collected and examined. Although these keywords were provided in Aaker's questionnaire to describe the brand, but Twitter users never know the Aaker's brand personality and personality dimension.

Chapter 4. Findings of the study

The purpose of Chapter 4 is to present the findings of Twitter data text mining based on the methodology implemented. Based on the methodology described in chapter 3, Twitter data collected from November 10 2014, to March 5 2015 were tested with Python text analysis scripts.

4.1. Hypothesis 1 and result

Hypothesis 1, Consumer Online Brand Related Activity will be detected on social media, was examined by collecting and cleaning data. The raw data included all tweets that include any keyword in the tweet. All of them were pushed from the Twitter HTTP streaming server and were analyzed with clustering, filtering, topic modeling, and part of speech tagging method.

	Apple	Samsung	iPhone	Galaxy	Total
Initial Strings	52,375,971	52,375,971	52,375,971	52,375,971	52,375,971
Cleaned Strings	1,419,793	1,656,565	1,493,478	1,502,449	6,072,285
Acceptance Level*	2.71%	3.16%	2.85%	2.87%	11.59%

Table 4.1 Number of tweets and its acceptance rate.

Acceptance Level* = $\frac{Number\ of\ Tweets\ After\ Data\ Cleaning}{Number\ of\ Raw\ Tweets}$

First of all, about 52.4 million tweets were collected and about 6.1 million of these tweets were used for text analysis. Language and topic filters strongly and rigorously excluded the tweets irrelevant to research topics. Those non-English and irrelevant topics were about 88.4% of total volume of tweet messages and about 11.6% of messages were English-based tweets relevant to study purpose.

Even though 88.4% of tweets were excluded by data cleaning, consumers on social media created or contributed fairly big amount of tweets about a brand and we could confirm that COBRA exists on social media; hypothesis 1 is supported. However, non-English tweets were excluded in this cleaned data set; but consumers using other languages also actively talking about a brand because Samsung and Apple product is sold globally. Thus, messages generated by COBRA might be larger than 11.6% of total tweet generated.

4.2. ADJECTIVE TERMS USED FOR KEYWORDS

After confirming that hypothesis 1, text analysis using natural language process was conducted to find the adjectives. Among these 6 million clean tweets, the volume of each keyword tweets were almost equal. Based on the cleaned tweets, adjective terms used to describe the keyword in each tweet were collected. However, rather than simple frequency of each terms, term's chance of appearance were calculated. In each keyword, how each adjective was used is following:

	Apple			Samsung		
Rank	Cleaned Strings	1,419,793	Chance of Appearance**	Cleaned Strings	1,656,565	Chance of Appearance
1	new	135,917	9.57%	new	139,362	8.41%
2	more	23,189	1.63%	more	26,160	1.58%
3	first	21,397	1.51%	first	22,992	1.39%
4	best	16,371	1.15%	big	19,080	1.15%
5	next	13,824	0.97%	best	17,946	1.08%
6	good	13,164	0.93%	good	16,674	1.01%
7	big	12,982	0.91%	hot	15,313	0.92%
8	real	12,863	0.91%	real	15,278	0.92%
9	green	12,695	0.89%	next	13,970	0.84%
10	mobile	12,013	0.85%	green	13,821	0.83%
11	own	11,723	0.83%	mobile	12,628	0.76%
12	electric	9,638	0.68%	own	12,489	0.75%
13	top	8,834	0.62%	top	11,993	0.72%
14	last	7,912	0.56%	electric	11,127	0.67%
15	old	7,896	0.56%	great	9,733	0.59%

Table 4.2 Adjective terms for Apple and Samsung

Chance of Appearance** = $\frac{Total\ number\ of\ strings\ which\ inclde\ the\ adjective}{Total\ number\ of\ strings\ after\ data\ cleaning}$

In keyword Apple and Samsung, most frequently used adjective was "new". The reason that "new" appeared the most is people purchased new Apple and Samsung products and people share it on Twitter. In addition to that, Samsung introduced their new flagship smartphone Samsung Galaxy S6 and Apple was planning to introduce the new wearable smart device, the Apple Watch, in the middle of data collection. Therefore, the adjective term "new" was used most to describe Apple and Samsung. Followed by "more", "first", "best", "good", "big", "real", "next", "same", and "free" were found as common adjectives for Samsung and Apple. 13 of 15 terms were commonly found to describe both brands; their brand personality adjectives on consumers are quite similar.

	iPhone			Galaxy		
Rank	Cleaned Strings	1,493,478	Chance of Appearance	Cleaned Strings	1,502,449	Chance of Appearance
1	new	134,221	8.99%	new	135,727	9.03%
2	more	25,563	1.71%	more	25,409	1.69%
3	first	21,785	1.46%	first	20,701	1.38%
4	best	15,725	1.05%	best	16,297	1.08%
5	good	15,505	1.04%	good	15,511	1.03%
6	next	14,045	0.94%	green	13,686	0.91%
7	green	13,752	0.92%	next	13,672	0.91%
8	big	13,632	0.91%	big	12,942	0.86%
9	real	12,922	0.87%	real	12,875	0.86%
10	own	12,265	0.82%	mobile	12,583	0.84%
11	mobile	12,095	0.81%	own	11,832	0.79%
12	electric	11,015	0.74%	electric	11,279	0.75%
13	top	9,573	0.64%	top	9,649	0.64%
14	bad	9,095	0.61%	bad	9,162	0.61%
15	last	8,461	0.57%	great	8,338	0.55%

Table 4.3 Adjective terms for iPhone and Galaxy

For iPhone and Galaxy, adjectives that describe the products in Twitter are almost similar. "New" appeared the most descriptive term used for iPhone and Galaxy; iPhone and Galaxy consumers described their experience with their "new" smartphone or they expressed their desire to purchase "new" smartphone. Followed by "more', "first", "best", "good", 14 out of 15 adjectives were commonly used to describe both iPhone and Galaxy.

Based on these text analyses, Hypotheses 3 and 3-1 were examined. As found in this research result, some adjectives were used more than others to describe each keyword; this result confirmed hypothesis 3. Therefore, as Biel (1992) argued, researcher could find some connection between attributes and image of brand in this research. Moreover, some generalizations of consumer's perception about two brands became possible because descriptive terms used were similar. Thus, brand personality could be studied by using social media.

However, lists of adjective terms describing Apple and Samsung, and adjective lists of iPhone and Galaxy were not different from each other. 13 out of top 15 adjectives describing Apple and Samsung brand were same; 14 out of top 15 adjectives describing iPhone and Galaxy were also found in both lists. Apple and Samsung brands described on Twitter were considered quite similar; this result indicates Apple and Samsung have at least similar brand personality. Thus, hypothesis 3-1 was not supported.

4.3. Aaker's brand personality adjectives on social media

In addition to text analysis to find brand personality adjectives, how often Aaker's 42 brand personality adjectives were used to describe Apple and Samsung brand in social media was tested by text mining. Similar to the Hypothesis 3 results, top 15 most frequently used Aaker's brand personality adjectives for Apple and Samsung were the same. Based on this result, it might be concluded that the Aaker's brand personality adjectives were also found in social media messages and might represent some brand personality of Apple and Samsung. Thus, hypothesis 4 was supported.

Rank	Apple		Chance of Appearance	Samsung		Chance of Appearance
1	real	12863	0.91%	real	15278	0.92%
2	original	2571	0.18%	young	4262	0.26%
3	cool	1934	0.14%	original	2200	0.13%
4	tough	761	0.05%	cool	2024	0.12%
5	young	745	0.05%	tough	781	0.05%
6	corporate	729	0.05%	corporate	736	0.04%
7	successful	446	0.03%	successful	506	0.03%
8	unique	359	0.03%	unique	369	0.02%
9	independent	330	0.02%	independent	327	0.02%
10	friendly	278	0.02%	friendly	285	0.02%
11	rugged	255	0.02%	technical	199	0.01%
12	technical	190	0.01%	rugged	195	0.01%
13	reliable	176	0.01%	reliable	169	0.01%
14	exciting	137	0.01%	exciting	146	0.01%
15	secure	129	0.01%	secure	127	0.01%

Table 4.4 Aaker's brand personality dimension adjectives in case of Apple and Samsung

Based on Aaker's brand personality adjectives result, I could confirm again that brand personality of Samsung and Apple was quite similar. Adjective "real" was most frequently used to describe Apple and Samsung followed by "original", "cool", and "tough", "young", etc.

In addition to Apple and Samsung, how Galaxy and iPhone were described by Aaker's adjectives was also tested; its result indicates that adjectives describing iPhone and Galaxy were quite similar, too.

Rank	Galaxy		Chance of Appearance	iPhone		Chance of Appearance
1	real	12875	0.86%	real	12922	0.86%
2	original	2146	0.14%	original	2607	0.17%
3	cool	1917	0.13%	cool	1942	0.13%
4	young	1155	0.08%	tough	768	0.05%
5	tough	767	0.05%	young	750	0.05%
6	corporate	723	0.05%	corporate	736	0.05%
7	successful	496	0.03%	successful	497	0.03%
8	unique	363	0.02%	unique	353	0.02%
9	independent	325	0.02%	independent	327	0.02%
10	friendly	283	0.02%	friendly	277	0.02%
11	rugged	195	0.01%	rugged	195	0.01%
12	technical	188	0.01%	technical	182	0.01%
13	reliable	169	0.01%	reliable	167	0.01%
14	exciting	126	0.01%	exciting	144	0.01%
15	honest	111	0.01%	secure	126	0.01%

Table 4.5 Aaker's brand personality dimension adjectives in case of Apple and Samsung

Repeatedly, adjective "real" was most frequently used to describe iPhone and galaxy followed by "original", "cool", and "tough", "young", etc. Interestingly, these Aaker's brand personality adjectives were used by social media users even the list of Aaker's brand personality adjectives were not provided. However, chance of appearances was less than 1% for all of keywords and only "real" was frequently used by social media users. In other words, Aaker's adjective terms to describe brand personality were not used well to describe Apple and Samsung brands, and iPhone and Galaxy. Perhaps, absence of the adjective list caused this low change of appearance. Also it can be argued that previous brand personality theory's survey's adjective terms were out-dated or could not reflect the brand personality of Apple and Samsung, smartphone makers. In addition to that, its order of most frequently used adjectives were quite similar because Aaker's brand personality adjectives are only 42 terms and their brand personalities are quite similar as discussed as findings. Thus, hypothesis 4-1 was not supported.

Chapter 5. General discussion

In Chapter 5, findings from this study, limitations, and future study plans will be discussed.

5.1 Discussion

Social media has become the new message channel for both brands and consumers. On social media, social media users share information about a brand and engage with a brand's activities and messages. These social media users' messages would include some descriptive terms about a brand and these descriptive terms would include and reflect the consumer's perception about a brand. The research questions were how consumers describe Apple and Samsung brand on social media and what extent these brand personalities were different. Even though only about 11.59% of tweets collected were usable for this text analysis, text mining and analysis results provide some insights about the Apple and Samsung brand.

The major finding of this research revealed that brand personalities were founded in Consumer Online Brand Related Activity in case of Apple and Samsung. Through data collection, data cleaning, and data analysis, hypotheses 1, 2, 3, and 4 were confirmed and supported. The findings have two huge implications; Consumer Online Brand Related Activity (COBRA) could be utilized for brand personality profile and future brand personality research application. Since Twitter users describe about their brand experience, thoughts, and feelings through their engagement with a brand and since there are millions of social media users and social media messages, some generalization about a brand becomes possible by collecting fairly large amounts of social media users' brand-related messages. As Plummer (1984) introduced in his paper, advertising and marketing practitioners utilize brand personality profile to check the consumer's perception of a brand. By collecting COBRA describing a brand, those advertising practitioners and marketers can check the consumer's perception of a brand.

On the other hand, researchers could examine brand personality theory with social media user's messages by text mining. In contrast to previous brand personality study's surveys, using social media data may provide larger sample than survey research. Moreover, rather than inflexible survey research in which questionnaires are fixed, research through text mining may provide more diverse opinions about brand personality. In this research, brand personalities of Apple and Samsung brands were examined by collecting and examining descriptive terms and by Aaker's 42 brand personality adjectives. Against hypotheses, results indicated that Apple and Samsung brands were quite similar and their flagship products also share similar images. However, this methodology can be utilized for future brand personality researches.

Additionally, this study may provide some implications for brand development and management. For example, a new brand may obtain some insights about another brand's personality from social media and apply it to their own brand personality. If a second tier brand wants to imitate the top tier brand personalities, social media data may provide insights about their targeted brand's personalities. On the other hand, a top tier brand may listen to what its followers express; it can differentiate further its brand personalities, image, products, services and brand communication strategies. Successful creation of a brand personality and image has great potential benefits; Esch, Langner, Schmitt, and Geus (2006) argue that brand image has a positive impact on brand trust, brand satisfaction and current purchases. Furthermore, brand trust has a positive impact on brand attachment and brand attachment has a positive impact on future purchases.

5.2 Limitations of the Study

This study has certain limitations. First of all, a powerful figure who/which has more followers in Twitter has more social influence to spread the word and cause an issue. For example, @UberFacts, the Twitter account for smartphone applications UberFacts promotion, has provided some interesting and fun information and it has 10.3 million followers. At January 29th, 2015, this account wrote message, “Apple could buy Disney and pay in cash”, and it is retweeted 2,604 times (Figure 5.1). Like UberFacts, there are thousands of accounts that have more than million followers and these figures have more social influence than other general social media users. and they can impact on follower’s product involvement and purchasing intention (Jin & Phua, 2014). In this study, these Twitter figures were excluded by best efforts because they are not general and ordinary consumers of Apple and Samsung brands.

Figure 5.1 Example of Twitter spike: UberFacts

Moreover, even though the researcher tried to filter out irrelevant data from raw data, there are chances that flaws are included in the data and may fail to sort out the irrelevant topics or noises. Since computer programs only conduct the analysis based on scripts and algorithms written by a human researcher, some irrelevant topics and data might be included in the text analysis data after data cleaning. For instance, apple in the text ‘An apple a day, keeps the doctor away’ and apple in ‘Apple to spend $19 billion on new European data centers’ will be considered the same by the computer, even though the first apple means the fruit apple and second apple is the brand Apple. In this study, both Latent Semantic Indexing and word collocation tested tweet collections relevant to research keywords. Using these methods, I could understand the usage and latent meaning of word Apple and Galaxy; then, irrelevant topics were filtered out through extra coding in this study. However, understanding, filtering, and sorting irrelevant keywords based on latent meaning were tedious and time-consuming work. Recently, in the academic discipline of computer science and engineering, machine learning, is widely discussed and studied. Machine learning is the area of study that a computer learns something from the data and makes decisions based on its learning. If through machine learning, a diverse meaning of Apple or Galaxy can be learned and sorted out, further study of brand or other keyword studies might be easier. Moreover, through machine learning, noise in social media might be understood and removed for brand analysis or consumer analysis.

Finally, Twitter has millions of users, but its users do not represent the whole consumers and population. According to pew research report, only 23% of Internet users in the United States use Twitter. Moreover, younger generation age 18 to 29 accounts for 37% of whole Twitter users; therefore, it is hard to say that Twitter users truly represent the United States consumers and population.

5.3 PLANS FOR FUTURE STUDY

In the next study, rival companies in other industries might be studied. For example, Coca-Cola and Pepsi or Toyota and Honda or Verizon and AT&T, are well known market rivals. Would they share a similar brand personality like the case of Apple and Samsung? If not, what extent are they different? Moreover, by testing other brands in other industries may confirm the brand personality theory or provide the counter example of theory.

For this study, only English tweets were collected and other languages were filtered. However, other languages should be eventually analyzed for better understanding of brands. Many brands and companies are advancing to almost every country on the earth as their products and services become available worldwide. Understanding global customers might provide another opportunity to strengthen the brand. There are many excellent language translators such as Google Translate and this study may be extended to another widely used language such as Chinese, Spanish, or Japanese to understand the global customer's perception of brand identity and brand image.

In summary, the findings of this research presented that brand Apple and Samsung, and iPhone and Galaxy were described almost similarly by social media users; Twitter users who were engaging with Apple and Twitter users who were engaging with Samsung have almost similar brand personalities and images in their mind, respectively. Also, even though social media seems promising source of big data, only 11.59% of data were usable for text analysis. This study provided a useful methodological insight for mining Twitter data and cleaning data for brand study. This thesis research will contribute to existing knowledge in the academic discipline of brand personality and social media and will offer valuable insights for effective strategic communication in social media.

References

Aaker, D. A., & Joachimsthaler, E. (2000). The brand relationship spectrum. *California Management Review, 42*(4), 8-23.

Aaker, J., & Fournier, S. (1995). A brand as a character, a partner and a person: three perspectives on the question of brand personality. *Advances in consumer research, 22*, 391-391.

Aaker, J. L. (1997). Dimensions of brand personality. *Journal of marketing research*, 347-356.

Aaker, J. L., Benet-Martinez, V., & Garolera, J. (2001). Consumption symbols as carriers of culture: A study of Japanese and Spanish brand personality constucts. *Journal of Personality and Social Psychology, 81*(3), 492.

Arndt, J. (1967). Role of product-related conversations in the diffusion of a new product. *Journal of marketing research*, 291-295.

Barnes, N. G., & Lescault, A. M. (2014). The 2014 Fortune 500 and Social Media: LinkedIn Dominates As Use of Newer Tools Explodes. from http://www.umassd.edu/cmr/socialmediaresearch/2014fortune500andsocialmedia/

Biel, A. L. (1992). How brand image drives brand equity. *Journal of Advertising Research, 32*(6), 6-12.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*: " O'Reilly Media, Inc.".

Bosnjak, M., Bochmann, V., & Hufschmidt, T. (2007). Dimensions of Brand Personality Attributions: A Person-Centric Aproach in the German Cultural Context. *Social Behavior and Personality: an international journal, 35*(3), 303-316.

boyd, d. m., & Ellison, N. B. (2007). Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication, 13*(1), 210-230.

Brandel, M. (1999). "1963: The debut of ASCII." Retrieved April 20, 2015, from http://edition.cnn.com/TECH/computing/9907/06/1963.idg/.

Brodie, R. J., Hollebeek, L. D., Juric, B., & Ilic, A. (2011). Customer engagement: conceptual domain, fundamental propositions, and implications for research. *Journal of Service Research*, 1094670511411703.

Caprara, G. V., Barbaranelli, C., & Guido, G. (2001). Brand personality: how to make the metaphor fit? *Journal of economic psychology, 22*(3), 377-395.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. *ICWSM, 10*(10-17), 30.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology, 37*(1), 51-89.

Chu, S.-C., & Kim, Y. (2011). Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites. *International Journal of Advertising, 30*(1), 47-75.

Corley, C. D., Cook, D. J., Mikler, A. R., & Singh, K. P. (2010). Text and structural data mining of influenza mentions in web and social media. *International journal of environmental research and public health, 7*(2), 596-615.

Crask, M. R., & Laskey, H. A. (1990). A positioning-based decision model for selecting advertising messages. *Journal of Advertising Research, 30*(4), 32-38.

De Longueville, B., Smith, R. S., & Luraschi, G. (2009). *Omg, from here, i can see the flames!: a use case of mining location based social networks to acquire spatio-temporal data on forest fires*. Paper presented at the Proceedings of the 2009 international workshop on location based social networks.

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., & Madden, M. (2015). Demographics of Key Social Networking Platforms *Pew Research Internet Project* (Vol. 30).

Duggan, M., & Smith, A. (2013). Social media update 2013. *Pew Internet and American Life Project*.

Esch, F.-R., Langner, T., Schmitt, B. H., & Geus, P. (2006). Are brands forever? How brand knowledge and relationships affect current and future purchases. *Journal of Product & Brand Management, 15*(2), 98-105.

Feick, L. F., Price, L. L., & Higie, R. A. (1986). People who use people: The other side of opinion leadership. *Advances in consumer research, 13*(1), 301-305.

Fleischman, M. (2014). The reach and impact of Oscars 2014 Tweets. Retrieved November 10th, 2014, from https://blog.twitter.com/2014/the-reach-and-impact-of-oscars-2014-tweets

Foux, G. (2006). Consumer-generated media: Get your customers involved. *Brand Strategy, 8*, 38-39.

galaxy. (n.d.). Dictionary.com Unabridged. Retrieved May 04, 2015, from Dictionary.com website: http://dictionary.reference.com/browse/galaxy

Gangadharbatla, H. (2012). Social Media and Advertising Theory. In S. Rodgers & E. Thorson (Eds.), *Advertising theory* (pp. 402-416): Routledge.

Geuens, M., Weijters, B., & De Wulf, K. (2009). A new measure of brand personality. *International Journal of Research in Marketing, 26*(2), 97-107.

Gilly, M. C., Graham, J. L., Wolfinbarger, M. F., & Yale, L. J. (1998). A dyadic study of interpersonal information search. *Journal of the Academy of Marketing Science, 26*(2), 83-100.

Google. (2013). Our Mobile Planet: United States of America

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management, 33*(3), 464-472.

Hearst, M. (2003). What Is Text Mining? Retrieved April 20, 2015, from http://people.ischool.berkeley.edu/~hearst/text-mining.html

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing, 18*(1), 38-52.

Interbrand. (2014). The Best Global Brand. Retrieved April 6th, 2015, from http://bestglobalbrands.com/2014/ranking/

Introducing JSON. (2015). Retrieved April 6th, 2015, from http://json.org

Jacobson, D., Woods, D., & Brail, G. (2011). *APIs: A strategy guide*: " O'Reilly Media, Inc.".

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology, 60*(11), 2169-2188.

Jin, S.-A. A., & Phua, J. (2014). Following celebrities' tweets about brands: The impact of twitter-based electronic word-of-mouth on consumers' source credibility perception, buying intention, and social identification with celebrities. *Journal of Advertising, 43*(2), 181-195.

Ko, H., Cho, C.-H., & Roberts, M. S. (2005). Internet uses and gratifications: A structural equation model of interactive advertising. *Journal of Advertising, 34*(2), 57-70.

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). *What is Twitter, a social network or a news media?* Paper presented at the Proceedings of the 19th international conference on World wide web.

Kyun Choi, Y., Kim, J., & McMillan, S. J. (2009). Motivators for the intention to use mobile TV: A comparison of South Korean males and females. *International Journal of Advertising, 28*(1), 147-167.

Langer, E. (2014). Why isn't @Apple on Twitter? *2015*(April 6th). Retrieved from: http://www.cnbc.com/id/101433491

Lannon, J., & Cooper, P. (1983). Humanistic advertising: a holistic cultural perspective. *International Journal of Advertising, 2*(3), 195-213.

Lasswell, H. D. (1948). The structure and function of communication in society. *The communication of ideas, 37*, 215-228.

Lawler, J. (1999). An English Word List. Retrieved March 2nd, 2015, from http://www-personal.umich.edu/~jlawler/wordlist.html

Levy, S. J. (1959). Symbols for sale. *Harvard business review, 37*(4), 117-124.

Lipsman, A., Mudd, G., Rich, M., & Bruich, S. (2012). The power of" like": How brands reach (and influence) fans through social-media marketing. *Journal of Advertising Research, 52*(1), 40.

Mangold, W. G., & Faulds, D. J. (2009). Social media: The new hybrid element of the promotion mix. *Business Horizons, 52*(4), 357-365.

McQuail, D. (2010). *McQuail's mass communication theory*: Sage publications.

Meraz, S. (2009). Is there an elite hold? Traditional media to social media agenda setting influence in blog networks. *Journal of Computer-Mediated Communication, 14*(3), 682-707.

Mostafa, M. M. (2013). More than words: Social networks' text mining for consumer brand sentiments. *Expert Systems with Applications, 40*(10), 4241-4251.

Muntinga, D. G., Moorman, M., & Smit, E. G. (2011). Introducing COBRAs: Exploring motivations for brand-related social media use. *International Journal of Advertising, 30*(1), 13-46.

Nandan, S. (2005). An exploration of the brand identity–brand image linkage: A communications perspective. *The Journal of Brand Management, 12*(4), 264-278.

O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM, 11*, 122-129.

Pak, A., & Paroubek, P. (2010). *Twitter as a Corpus for Sentiment Analysis and Opinion Mining*. Paper presented at the LREC.

Plummer, J. T. (1984). How personality makes a difference. *Journal of Advertising Research, 24*(6), 27-31.

Powers, T., Advincula, D., Austin, M. S., Graiko, S., & Snyder, J. (2012). Digital and Social Media In the Purchase Decision Process A Special Report from the Advertising Research Foundation. *Journal of Advertising Research, 52*(4), 479-489.

Řehůřek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora.

Richins, M. L. (1984). WORD OF MOUTH COMMUNICATION A NEGATIVE INFORMATION. *Advances in consumer research, 11*(1).

Roesslein, J. (2014). Tweepy (Version 2.3.0). Retrieved from http://www.tweepy.org

Ruggiero, T. E. (2000). Uses and gratifications theory in the 21st century. *Mass communication & society, 3*(1), 3-37.

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors*. Paper presented at the Proceedings of the 19th international conference on World wide web.

SAS Institute. (2010). Social Media Metrics: Listening, Understanding and Predicting the Impacts of Social Media on Your Business.

SIL International. (2005). English wordlist. Retrieved March 2nd, 2015, from http://www-01.sil.org/linguistics/wordlists/english/

Singletary, T. (2012). 1% of the firehose via streaming API even when only searching very specific keywords? Retrieved April 6th, 2015, from https://twittercommunity.com/t/1-of-the-firehose-via-streaming-api-even-when-only-searching-very-specific-keywords/9712

Smith, A. (2013). Smartphone Ownership 2013: Pew Research Center.

Solis, B. (2010). *Engage: The complete guide for brands and businesses to build, cultivate, and measure success in the new web*: John Wiley & Sons.

Sun, T., Youn, S., Wu, G., & Kuntaraporn, M. (2006). Online word-of-mouth (or mouse): An exploration of its antecedents and consequences. *Journal of Computer-Mediated Communication, 11*(4), 1104-1127.

Thorson, E., & Rodgers, S. (2012). What Does "Theories of Advertisng" Means. In S. Rodgers & E. Thorson (Eds.), *Advertising theory* (pp. 3-17): Routledge.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. *ICWSM, 10*, 178-185.

Twitter Inc. (2015a). API Rate Limits. Retrieved April 6th, 2015, from https://dev.twitter.com/rest/public/rate-limiting

Twitter Inc. (2015c). Documentation. Retrieved April 6th, 2015, from https://dev.twitter.com/overview/documentation

Twitter Inc. (2015e). GET help/languages. Retrieved April 6th, 2015, from https://dev.twitter.com/rest/reference/get/help/languages

Twitter Inc. (2015g). Processing streaming data. Retrieved April 6th, 2015, from https://dev.twitter.com/streaming/overview/processing

Twitter Inc. (2015i). REST APIs. Retrieved April 6th, 2015, from https://dev.twitter.com/rest/public

Twitter Inc. (2015k). Streaming API request parameters. Retrieved April 6th, 2015, from https://dev.twitter.com/streaming/overview/request-parameters

Twitter Inc. (2015n). The Streaming APIs. Retrieved April 6th, 2015, from https://dev.twitter.com/streaming/overview

Twitter Inc. (2015p). Twitter Libraries. Retrieved April 6th, 2015, from https://dev.twitter.com/overview/api/twitter-libraries

Twitter Inc. (2015r). Working with Timelines. Retrieved April 6th, 2015, from https://dev.twitter.com/rest/public/timelines

Van-Tien Dao, W., Nhat Hanh Le, A., Ming-Sung Cheng, J., & Chao Chen, D. (2014). Social media advertising value: The case of transitional economies in Southeast Asia. *International Journal of Advertising, 33*(2), 271-294.

Vickery, G., & Wunsch-Vincent, S. (2007). *Participative web and user-created content: Web 2.0 wikis and social networking*: Organization for Economic Cooperation and Development (OECD).

Vollmer, C., & Precourt, G. (2008). *Always on: Advertising, marketing, and media in an era of consumer control*: McGraw Hill Professional.

Wang, X., Yu, C., & Wei, Y. (2012). Social media peer communication and impacts on purchase intentions: A consumer socialization framework. *Journal of Interactive Marketing, 26*(4), 198-208.