TOWSON UNIVERSITY

COLLEGE OF GRADUATE STUDIES AND RESEARCH

A STUDY OF ASPECT-BASED SENTIMENT ANALYSIS IN SOCIAL MEDIA

by

Youngsub Han

A Dissertation

Presented to the faculty of

Towson University

in partial fulfillment

of the requirements for the degree

Doctor of Science in Information Technology

Department of Computer and Information Sciences

Towson University

Towson, Maryland 21252

May 2017

TOWSON UNIVERSITY

OFFICE OF GRADUATE STUDIES

DISSERTATION APPROVAL PAGE

This is to certify that the thesis prepared by Youngsub Han, entitled "A Study of Aspect-Based Sentiment Analysis in Social Media," has been approved by the thesis committee as satisfactorily completing the dissertation requirements for the degree of Doctor of Science in Information Technology.

Chairperson, Dissertation Committee, Dr. Yanggon Kim — Date: 4-19-2017

Committee Member, Dr. Sungchul Hong — Date: 4/19/2017

Committee Member, Dr. Michael P. McGuire — Date: 4/19/2017

Committee Member, Dr. Siddharth Kaza — Date: 4/19/2017

Dr. Barin Nag — Date: 4/19/2017

Janet V. DeHaney, Dean Graduate Studies — April 27, 2017

Acknowledgement

I would like to gratefully thank my advisor, Dr. Yanggon Kim, for his guidance, caring, patience, and providing me with an excellent atmosphere for doing research. I would like to thank Dr. Kwangmi Kim for her assistance and guidance of my research. In addition, I would like to thank my committee members, Dr. Sungchul Hong, Siddharth Kaza and Michael McGuire for their input, valuable discussions and accessibility. I would like to sincerely thank colleagues, Hyeoncheol Lee and Beomseok Hong. My completion of this dissertation could not have been accomplished without them.

I would like to express deepest gratitude to my parents, Yoo-Jeon Han and Young-Hee Choi, and my wife's parents Yong-Cheol Kim and Sang-Sook Lee. Especially, my brother, Jung-sub Han, who is my first teacher on program language and put me on this journey. They were always supporting me and encouraging me with their best wishes.

Finally, and most importantly, I would like to attribute all my achievement to my beloved wife, Yuyoung Kim. Her constant support, patience, tolerance, sacrifices and love allowed me to undertake this research.

Above all, I would like to thank God, the almighty, for giving me the wisdom, knowledge, strength, opportunity and everlasting love.

ABSTRACT

A Study of Social Media Analysis using Aspect-Based Approach

Youngsub Han

In recent years, diverse social media platforms, such as Facebook, YouTube, Instagram, Snapchat, and Twitter, have rapidly grown in size and influence. Various industries, including the media and advertising industries, have made significant efforts to adapt to building competitive advantages using social media in the marketplace. Therefore, it became imperative for marketing and advertising professionals to better understand the complex behaviors and minds of consumers using data-mining techniques, which help to handle massive amounts of social media data.

In this research, we analyzed Twitter to discover characteristics of social media. This study is intended to address these topics to build a better understanding of Twitter usages. Even if there are many important theories and frameworks in media studies, we were not able to single out one specific theoretical framework for this study. Therefore, by using the active audience concept, and relying on marketing literature, we chose a grounded theory approach and presented research questions for in-depth understanding of Twitter usage in order to detect any patterns that consumers might show.

In addition, we proposed a sentiment analysis method and a lexicon building method using the morphological sentence pattern (MSP) model. These methods aim to observe and summarize people's opinions or emotional states from the social media data. Despite the demands of sentiment analysis methods for analyzing social media data, fundamental challenges remain because user-generated online textual data is unstructured, unlabeled, and noisy. Finally, we proposed a method of extracting movie genre similarity for movie recommendation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

In recent years, we have witnessed a phenomenal change in how we communicate, largely due to the explosion of social media and technologies. Diverse social media platforms, such as Facebook, YouTube, Instagram, Snapchat, and Twitter, have rapidly grown in size and influence. Particularly, Twitter has become one of the popular social media platforms since its inception in 2006, with over 300 million users. Recently, the Pew Research Center reported that Twitter users account for about 23% of all Internet users, and 63% of adults (ages between 18 and 49) use Twitter [1]. Since the first tweet was sent by Jack Dorsey (@jack) on March 21, 2006, the average number of tweets per day has increased from 300,000 in 2008 to 500 million in 2017 [2].

Traditionally, the flow of mass communication has been unidirectional, originating from businesses and organizations to inform, persuade, and remind current and potential customers of their product offerings and benefits particularly in advertising. In this traditional model, consumers were passive receivers and simply reacted to such messages by either ignoring or becoming attentive [3, 4, 5]. However, technology and the Internet have revolutionized this traditional communication pattern and have transformed it into a more interactive process. This phenomenon allows consumers to be more connected, informed, and empowered in this social media age.

Various industries, including the media and advertising industries, have made significant efforts to adapt to such swift changes and have paid attention to building competitive advantages using social media in the marketplace. Social media has become an important marketing venue, allowing them to reach a wide range of target audiences efficiently. Therefore, it is typical to see icons of Facebook and Twitter on many businesses' websites and advertising messages. Simultaneously, the same social media sites create serious challenges for the marketing world. Since consumers are further fragmented by various media platforms, they can jeopardize brand equity and brand images by sharing their unpleasant or dissatisfied experiences with others through social media [6]. Therefore, it became imperative for marketing and advertising professionals to better understand the complex behaviors and minds of consumers using data-mining techniques, which help to handle massive amounts of social media data and comprehend the nature of social media data exchanged for a certain brand or commercial instead of using a traditional content analysis.

First, we analyzed Twitter to discover characteristics of social media with research questions: "Why do consumers choose to use Twitter? How do they use it? What interactions do they have with other users?" This study is intended to address these topics to build a better understanding of Twitter usages. There are many important theories and frameworks in media studies; we were not able to single out one specific theoretical framework for this study. Therefore, by using the "active audience concept," and relying on marketing literature, we chose a grounded theory approach and presented research

questions for in-depth understanding of Twitter usage in order to detect any patterns that consumers might show.

Second, we proposed a sentiment analysis method using the morphological sentence pattern (MSP) model [10, 11]. This method aims to observe and summarize people's opinions or emotional states from textual data. Despite the demands of sentiment analysis methods for analyzing social media data, fundamental challenges remain because user-generated online textual data is unstructured, unlabeled, and noisy. In addition, we proposed a lexicon building method for the sentiment analysis method because the lexicon building usually needs human-coding efforts to maintain quality of analysis [7, 8, 9].

Third, movie recommendation is a field of advertisement or promotion that target customers in the movie industry. Accordingly, various companies use recommendation systems for the success of their business. In this research, we propose an extracting method of movie genre similarity for movie recommendation using the MSP model and machine learning algorithms. In this method, we proposed two main methods, which are "TDF-IDF" and "Genre Score." The TDF-IDF is designed to find genre representative keywords and the Genre Score is designed to discover a similarity of movies with consideration for genres using the keywords. The results of these methods are used as features to find similar movies. In addition, we used machine-learning algorithms such as K-Means and K-Nearest Neighbor (KNN) algorithms to recommend movies.

We describe our work as follows in-detail at the rest of sections. In Section 2, we present literature reviews related to our research. In Section 3, we present a system architecture, which is designed to handle social media data in real-time. In Section 4, we present a study of analyzing Twitter-usage-patterns and its characteristics based on research questions. In Section 5, we propose a sentiment analysis method and a lexicon building method for extracting useful information from the social media data. In Section 6, we propose a method for movie recommendation. In Section 8, we conclude this study with summarization and contributions.

2. LITERATURE REVIEW

2.1. Social Media Analysis

2.1.1. Active Audience and Uses and Gratifications

A common assumption held by social psychologists and technology adoption researchers is that few media can fulfill all the goals audiences seek. Accordingly, the audiences select certain media based on their perceived functionality and use several media at the same time [12, 13]. Recent industry reports indicate that consumers mix traditional and digital media at the same time, and TV and online media was the most popular combination. For example, eMarketers' survey conducted in December 2011 reported that people used TV for 3.4 hours and online media for 3.1 hours while the Interactive Advertising Bureau reported TV using time for 4.6 hours and Internet for 2.8 hours in 2012 [14]. The most recent industry data, however, reported that TV consumption dominance over Internet was reversed in 2013 as people spent more time with digital media than watching TV, and projected that people will spend about 6 hours on digital media and about 4 hours on TV by 2017 [15]. In addition, more than half of all media interactions involve multitasking, and about 77% of people use TV with another device. In particular, about 49% of media users use TV with a smartphone and about 35% use TV with a PC/laptop [14]. These changing media consumption behaviors have made marketers and various organizations diversify the channels to reach their target audiences.

As an influential theory in media research, the Uses and Gratification (U&G) perspective assumes that people can use the same medium for different purposes. The theory holds that multiple forms of media compete for users' attention and audiences select the medium that meets their needs, such as the desire for information, emotional connection, or status [16]. At the core of this theory is the concept of an active audience, which assumes that the audience's communication behavior is goal-oriented and purposeful in that people choose certain media based on their needs, wants, or expectations. U&G has recently been revitalized for studying technologies and media consumption behavior. This includes research on the web [17], on blogging [18], and social-networking sites, such as Facebook and Twitter [19, 20]. Researchers found that interactivity, recreation, entertainment, diversion, information involvement, connectedness, and personal relevance are all major motivating factors to browse or use the Internet and social media platforms.

Particularly, Stafford, Stafford and Schkade [21] have identified that users seek three types of gratifications: content gratification (the content of the medium, whether it's entertainment or information), process gratification (the experience of the media usage itself, such as Internet surfing or experiencing a new technology), and social gratification (the interpersonal communication and social networking opportunities on the Internet). Shao [22] further argues that individuals use online media at three different stages/levels for their own needs: (1) consuming content for information, entertainment, and mood management needs; (2) participating through interacting with the content, as well as with

other users, for social connections; and (3) producing their own content for self-expression and self-actualization. In other words, online users have varying degrees of engagement with social media, ranging from simple and passive (e.g., simply consuming the contents by reading) to active (e.g., producing and posting the contents).

What Shao [22] indicates in his study has relevant implications for the analysis of the different message types that people tweet. The first level -- consuming content for information, entertainment, or mood management needs -- indicates a simple, passive reading behavior of users. On the other hand, the second and the third levels involve more active roles of users from tweeting their own thoughts, emotions, and information to retweeting others' messages and further to replying to certain messages for a higher level of engagement and social interactions. Based on this implication, one of this study's objectives is to identify dominant or popular types of tweets that people use.

2.1.2. Social Media and Marketing

No one can dispute that social media have become a vital marketing tool in the twenty-first century, particularly among millennials. As consumers become active in expressing their opinions about brands through reviews, microblogs, pictures, and video blogs, marketers have made more conscious efforts to engage consumers in building relationships. One example is what Doritos has done over the past ten years by running consumer-generated commercials contest for the Super Bowl games. Doritos was able to

engage the consumers and increase the consumer loyalties [23]. Due to the importance of social media in marketing, particularly in relationship marketing, consumers as well as marketers acknowledged that social media should be part of overall brand communications. In 2009, 16% of the Fortune 500 companies had corporate blogs that link to a variety of social media channels, including podcasts, RSS feeds, and Twitter [24]. In 2013, 86% of marketers believe social media is an important channel for their marketing initiatives [25]. In addition, about 93 % of consumers indicated that a company should have a presence in social media while 85% indicated that a company should seek active interactions with customers through these platforms [26]. Such recognition of the importance of social media as a marketing tool was reflected in their increased social media spend. Total spend on social media advertising has increased by 56.2% from $11.36 billion in 2013 to $17.74 billion in 2014 [27].

Among the growing number of studies on social media and marketing, some literature has focused on consumers' responses to consumer-generated advertising (CGA), or user-generated content (UGC) and examined whether they differ from firm-generated advertising (FGA) or firm-generated content (FGC) [28, 29, 30]. Pehlivan et al. [30] analyzed consumers' comments left for CGAs and FGAs for the MacBook Air, and found that the nature of comments for each type of ads was different. Comments for CGAs were more focused on humor while comments for FGAs were referenced to the major features, such as the song used in the ads. Another study in a similar topic examined the effects of FGCs in social media and found that FGCs not only enhance the

transaction and relationship sides of customer-firm interactions but also play a role in increasing profitability [29]. These researchers also found that FGCs became more effective when used simultaneously with other communication channels, such as TV and emails. These studies guided another research question for this study and led us to examine the sources of tweets whether they were generated by individual personal consumers (similar to the concept of user-generated contents or consumer-generated ads) or by commercial business-oriented sources (similar to the concept of firm-generated ads or contents). As Twitter becomes a more vital marketing tool, many companies and websites have recently provided Twitter-based advertising services and business solutions. Those companies, such as Unfollowers, TweetDeck and TweetAdder, tend to generate tweets automatically to enhance certain brands' performances in the marketplace. Considering this growing trend in the Twitter industry, we also would like to see how strong these activities are in section 4 with experiment results based on the research questions.

2.1.3. Social TV

Schirra, Sun, and Bentley [34] provided in-depth analysis of live-tweeting motivations by interviewing 11 tweet users on "Downton Abbey," a PBS serial drama broadcast in 2013. In this study, they revealed that there are some live-tweeting "triggers" that prompted TV viewers to post tweets about a show. They included various emotions

that the viewers felt while watching TV programs (e.g., sadness and grief), comedic moments of the program (e.g., humor) and character developments within the program. They also noted that people tweet during TV viewing for various personal benefits, such as gaining a sense of connectedness with a broader audience to avoid loneliness or to confirm their thoughts or opinions. Therefore, the various and complex individual and social motivations of using social media as people watch TV revealed various levels of communication activities by users. Some people simply post messages for self-expression and self-actualization while others have strong desire to maintain social interaction as one of major motivations for media consumption [13, 31].

Overall, Twitter as one of the most popular social media platforms is expected to enhance social interactions among remotely located viewers and even to form a "community" through shared viewing experiences. At that same time, literature on Twitter use reports that the interactivity among Twitter users tend to be lower than expected as less than four percent of tweets were interactive during TV viewing [32]. Based on this mixed understanding, this study seeks to examine the level of activity among TV viewers – whether people who post messages on Twitter are just posting their thoughts and emotions without strong desire to connect with others, or actually sharing/interacting with other people.

Previous work related to social TV revealed the effect of different genres on social TV activities. Wohn and Na [32] analyzed and compared the content of tweets

posted about two different types of TV programs: One was a political speech by President Barack Obama at the White House announcing his acceptance of the Nobel Peace prize and the other was a reality TV show, an episode of ABC network's "So You Think You Can Dance (SYTYCD)." Categorizing the messages in an attention, emotion, information, and opinion (AEIO) matrix, Wohn and Na [32] found that the content of tweets correlated with a TV program genre. For example, attention messages for SYTYCD were noticeably higher than those for Obama speech, and the opinion message type was the most salient for Obama's speech. In addition, tweet messages peaked right after President Obama made his acceptance of the award. For SYTYCD, they revealed that information message type was higher before the beginning of the program, followed by attention messages as viewers were about to watch the program. They also observed a rise in tweeting activity during commercial breaks.

Similarly, Doughty, Rowland, and Lawson [35] reported that different types of TV programs yielded different communication activities. They analyzed tweet messages about two different UK TV shows: "The X Factor" and "BBC Question Time." The X Factor was a prime time, TV talent show with entertainment elements, while "BBC Question Time" is a late night, current affairs panel debate program with heavy informational elements. Doughty et al found that "The X Factor" generated much shorter 'outburst' type of messages in line with entertainment nature of the show, and "The Question Time" generated longer tweet messages as the audiences contributed to the political debate and discussion. Another interesting finding was that there were no

differences in messages by different platforms (e.g., mobile devices and web clients). In other words, they suggested that TV shows themselves are the leading factor in shaping the message contents.

Confirming the general literature finding on this topic, Buschow, Schneider, and Simon [33] also revealed that three different TV genres in Germany evoked different communication activities. They analyzed two TV shows for each of three categories. The first category, talent shows, produced expressions of freedom and critiques of the candidates in the show, and the second category, live events broadcast, evoked a critical debate about the show itself and the event as it evolves on screen. The third program category, political debate shows, stimulated a public discourse.

Social network sites allow television viewers to enjoy the communal experience of group viewing without physically being together. Viewers share their viewing experiences real-time through computer-mediated communication, which creates a pseudo-communal viewing experience. Typically, social media is well known for assisting this new form of TV viewing practice [32, 33]. Accordingly, we now have several terms that are related to this changed communication consumption behavior, such as “real-time backchannel conversations,” “digital backchannels,” “live-tweeting” or “social TV.” [33, 34, 35, 36]

There is an emerging body of literature on what types of messages people share with others while they are watching TV and how those messages and conversations are

related to the context of the program they are watching [32, 33, 34, 37]. However, little research has been conducted on social TV with a sport game. Therefore, this study plans to analyze the viewer's social TV behaviors and engagement during a Super Bowl game and further compare whether the nature of the game is related to the "coverage of conversations" in section 4.

2.2. Sentiment Analysis in Social Media

The purpose of sentiment analysis is to discover and summarize opinion or emotional states regarding certain topics such as events, products, entertainers, politicians, movies, services from the textual data to find people's interesting and thought [9]. In this section, we describe existing approaches of sentiment analysis and related methods.

Guerra et al. proposed a sentiment analysis algorithm to measure the bias of social media users toward a topic [41]. They hypothesized that users tend to express their opinion multiple times and a user's bias tends to be more consistent over time as a basic nature of human behavior. Thus, the bias toward a topic of social media users was measured and the sentiment was analyzed by transferring user's biases into textual features.

Kucuktunc et al. also proposed a method to analyze sentiment based on demographic characteristics of users, such as gender, age and education [42]. However, these methods cannot be broadly used because they required users' relationship data and previous messages that the users posted, which are not always available in social networks due to the privacy laws.

Speriosu et al. applied a label propagation (LPROP) approach based on graph representation to analyze the sentiment of messages in Twitter [13]. Their assumption was that each tweet written by a user is linked to other tweets written by the same user, and each user is influenced by tweets written by other users that he or she follows. They represented such a relationship using a graph where the features of the message, such as words, emoticon and authors, are inter-related to each other. Those features determine the positivity or negativity of the message in the graph. In addition, the results were examined based on the accuracy of the LPROP approach with messages in four different topics. The accuracy of the proposed LPROP approach is the highest among other sentiment analysis approaches as its level ranges from 65.7% to 84.7%, depending on the topics. However, we think this accuracy could be further improved because its average accuracy is still 72.08%.

Lexicon-based approach is a traditional sentiment analysis approach. O'Connor et al. analyzed political opinions using this approach [44]. They collected tweets related to political issues from 2008 to 2009. Then, they built a lexicon where each word was

categorized as either positive or negative keywords based on OpinionFinder [45]. Once positive and negative keywords were counted for all the messages, each message is classified as positive or negative. As a result, the ratio of positive messages versus negative messages was compared with survey results, which showed a strong correlation (80%) between the results of the sentiment analysis and survey results. The finding suggested that the lexicon-based method could be used as a supplement for traditional survey. However, this lexicon-based approach also has a weakness in that a message including positive keywords does not necessarily yield a positive opinion all the time. For instance, a word "like" is categorized as a positive word in the lexicon, and hence, if a message includes the word "like," it is categorized as a positive message. However, if the message includes the word "don't" right before "like," the actual opinion of message should be categorized as negative. Such limitation should be improved with lexicon-based approach.

In our previous research, we proposed a method for sentiment analysis using the probability model [46]. The method requires a train-set with a human-coded process and builds a sentiment lexicon that contains the list of words that appeared in the text messages. For example, when a word occurs in seven positive messages out of 10 messages, the word would have 0.7 (70%, 7/10) probability. Then, it computes the positivity score of text messages in a test-set using the list of words in a message and sentiment lexicon. Each message is categorized as either positive or negative, depending on threshold value calculated using a train set. As we mentioned, this model also has

limitations as all words can have positivity even though words are not meaningful. It causes over-analysis to extract opinion from the data. In addition, the train-set must be updated frequently as online data and messages change.

The aspect-based sentiment analysis is a lexicon-based approach because this approach uses the lexicon as a measurement. The major difference between the two is that the aspect-based analysis provides a more in-depth analysis because all results are categorized into each aspect. In this approach, an aspect seems an attribute of objects. For example, when an object is a mobile phone, its aspects can be “display,” “size,” “price,” “camera,” or “battery” to describe the mobile phone. Thus, expected results are paired with an aspect and an expression such as “display-clean”, “price-good”, or “camera-awesome” [47, 48, 49]. Therefore, we will adopt this approach for more in-depth analysis.

In lexicon-based sentiment analysis, the lexicon building is a fundamental challenge because the lexicon is a main measurement for extracting opinions from the data and it usually requires human-coding efforts. For example, when lexicon contains low polarity words or meaningless words, these may cause a bad influence on the accuracy. Many studies have proposed unsupervised or semi-supervised approaches for building lexicon. These recent studies have focused on constructing emotional lexicon that assign into fine-grained categories of emotions, such as happiness, like, disgust, sadness, and anger, [50, 51, 52, 53, 54].

Particularly, J. Bross, and H. Ehrig proposed a method that allows to existing lexicon to adapt and extend automatically for a specific product domain [47]. However, they simply used morphological patterns to extract aspects and expressions. Therefore, we propose a method to extract aspect-based lexicon using a morphological sentence pattern model as well. In Section 5, we describe the proposed method in-detail.

In rest of this section, we describe three methods related to propose our methods. First, Natural Language Processing (NLP) is an area of computer science to understand human languages. NLP is widely used as an essential process to analyze textual data in data mining, artificial intelligence, and computational linguistics speech recognition [63]. User-generated textual data in social media data have many linguistic problems such as spacing errors, idioms, and jargons. Accordingly, we use a natural language processing tool which is the "Stanford Core NLP" made by The Stanford Natural Language Processing Group [10]. This tool provides refined and sophisticated results based on English grammar [40]. In section 5, we will describe how this tool was applied to our method in-detail.

Second, Term Frequency-Inverse Document Frequency (TF-IDF) is widely used to extract keywords from textual data. This algorithm was designed for calculating the importance of keywords in corpus. The base form of the equation is:

$$tfidf_{t,d} = (1 + log(tf_{t,d})) \times (log\left(\frac{N}{df_t}\right)) \quad (1)$$

In this method, TF (Term Frequency) is a measurement of how frequently a keyword occurs in a document, and IDF (Inverse Document Frequency) is how each keyword is less-commonly used in the corpus. IDF is the number of documents divided by the document's frequency of a keyword in all documents. Therefore, the TF-IDF is TF multiplied by the IDF and it indicates a degree of document specific keywords in corpus. In this research, we used a normalized TF-IDF using the logarithmically scaled frequency to prevent a bias towards longer documents [55, 56]. In section 6, we proposed a "TDF-IDF" based on this algorithm.

Third, F-measure (F-score) is broadly used to measure the performance for sentiment analysis systems [57]. The definition of the measure is:

$$\mathrm{F-measure} = 2 * \frac{\mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{2}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

This algorithm is a measure of a test's accuracy with considering both "Precision" and "Recall" (2). The recall means the portion of relevant instances that are retrieved (3), and the precision means the portion of retrieved instances that are relevant (4). In these equations, TP is true positive, which indicates a proportion of keywords correctly labeled as belonging to the answer-set. FP is false positive, which indicates a proportion of

extracted keywords incorrectly labeled as belonging to the answer-set. FN is false negative, which indicates a proportion of extracted keywords are not labeled as belonging to the answer-set. Using this method, we compered quality of our methods with existing methods.

2.3. Movie Recommendation and Algorithms

In 2015, filmmakers launched 2,087 movies around the world. This number is a 26% growth from 2014 according to the theatrical market statistics in 2015 released by Motion Picture Association of America (MPAA, http://www.mpaa.org, Theatrical market statistics 2015). Moreover, online video streaming is quickly growing through streaming services such as Netflix, Hulu Plus, and Amazon. Accordingly, many studies have been conducted to recommend movies because a movie recommendation is a field of advertisement or promotion that target customers in the movie industry [94, 95].

A recommendation system is one of the information filtering techniques to indicate information items such as movies, music, web sites, news that are likely of interest to the customers. Various companies such as Netflix, Google and Amazon use recommendation systems for the success of their business. To recommend movies, various information can be used as features such as movie reviews, comments, view histories, ratings, actors, producers, writers, running time, box office, and genres [67, 95,

96, 97]. In this research, we focused on the movie reviews and YouTube comments to discover characteristics of movies to consider people's perspectives because user-generated data contain their opinions including unpleasant or dissatisfied experiences [6]. Text-based approaches are commonly used on video classification [98]. The advantages of text-based approaches are the utilization of a large body of research conducted on text classification, which helps to understand the user's perspectives [98, 99].

Orellana-Rodriguez et al. proposed a text-based approach that automatically extracts affective context from user comments associated with short films available on YouTube to explore the role of an emotional state as an alternative to explicit human annotations. All extracted emotional keywords would be categorized into four representative categories: joy-sadness, anger-fear, trust-disgust, and anticipation-surprise. The results showed how the affective context could be influenced for emotion-aware film recommendation [96]. Diao et al. proposed a movie recommendation model using an aspect-based approach from online movie reviews and ratings [97]. The aspect-based approach provides more in-depth analysis than the traditional lexicon-based approach [47, 48, 49].

K-Means clustering algorithm and K-Nearest Neighbor (KNN) algorithm are used to recommend relevant movies in this research. Clustering is an unsupervised learning algorithm for dividing a set of objects into smaller sets [58]. K-Means clustering algorithm was originally proposed in 1965 by Forgy [59] and in 1967 by MacQueen [60].

This algorithm is still one of the most popular clustering algorithms in data mining, artificial intelligence and computer graphics. This algorithm aims to partition *n* observations into *k* clusters. Each observation belongs to the cluster with the nearest mean in the cluster [61]. Two main features are the Euclidean distance as a metric for measuring the distances between the points, and the number of clusters *k*, given as an input parameter to the algorithm [62].

In addition, Elbow method is used to determine the number of clusters (*k*) [62]. This is a method of interpretation and validation of consistency within cluster analysis designed to find the appropriate number of clusters (*k*). In the Elbow method, the percentage of variance is explained as a function of the number of clusters. A number should choose a number of clusters so that adding another cluster does not give much better modeling of the data. More precisely, if one plots the percentage of variance explained by the clusters against the number of clusters, the first clusters will add much information such as variances, but at some point the marginal gain will drop, giving an angle in the graph. The number of clusters is chosen at this point [62].

K-Nearest Neighbor (KNN) algorithm is used for classification and regression [64]. In classification, this method classifies an object by a majority vote of its neighbors. Euclidean distance is commonly used as a metric to determine their similarity. If $k = 1$, then the object is simply assigned to the class of that the nearest neighbor [65, 66]. T. Loren et al. used KNN method for collaborative filtering, which is a filtering process for

information or patterns on collaboration of objects such as users or data sources on recommender system [67]. Thus, we used these algorithms to find relevant movies based on their similarity to recommend movies (see chapter 6).

3. SYSTEM ARCHITECTURE

A huge amount of data is being generated though social media in real-time. Many researchers have developed data collection and management systems for online social network analysis, which is summarized in Table 1. Most of the previous research encompasses potential problems with data processing, management and analysis. Firstly, the data storages of the previous approaches are based on a relational database that may cause performance issues when a huge amount of datasets ranging from a few terabytes to multiple petabytes needs to be handled. Secondly, they do not support distributed processing, which may slow down processing time. Lastly, they collect data from only a single source channel, such as Twitter. To analyze trends of society accurately, data should be collected from multiple online social networks as opposed to only one.

Table 1. Comparison of related approaches

	Source Channel	**Data Store**	**Distributed Processing**
Song et al [112]	Twitter	Relational, Key-value pairs	No
TwitterEcho [113]	Twitter	Not given	No
Byun et al [114]	Twitter	Relational	No
Twitter Zombie [115]	Twitter	Relational	No
TwitHoard [116]	Twitter	Graph DB	No
TrendMiner [117]	Twitter	Key-value pairs	No
TwitIE [118]	Twitter	Not given	No
ESA [119]	Twitter	Not given	No
Baldwin et al [120]	Twitter	Flat files	No

For these reasons, we propose a system to discover meaningful information such as keywords and emotional states from the social media data. Figure 1 shows the system architecture and flow. This system consists of four main phases, data collector, pre-processing, analyzer, and data controller. In addition, this system is designed to support multi-processing for reducing the gap in performance using distributed architectures. Particularly, we used ActiveMQ for communicating data between phases. The ActiveMQ is an open-source message broker written in Java with Java Messaging Service (JMS). We also used Hadoop (HDFS) as a main data repository. For a convenience of accessing data, we used a relational database system, MySQL. In the rest of this section, we describe each phase in-detail.

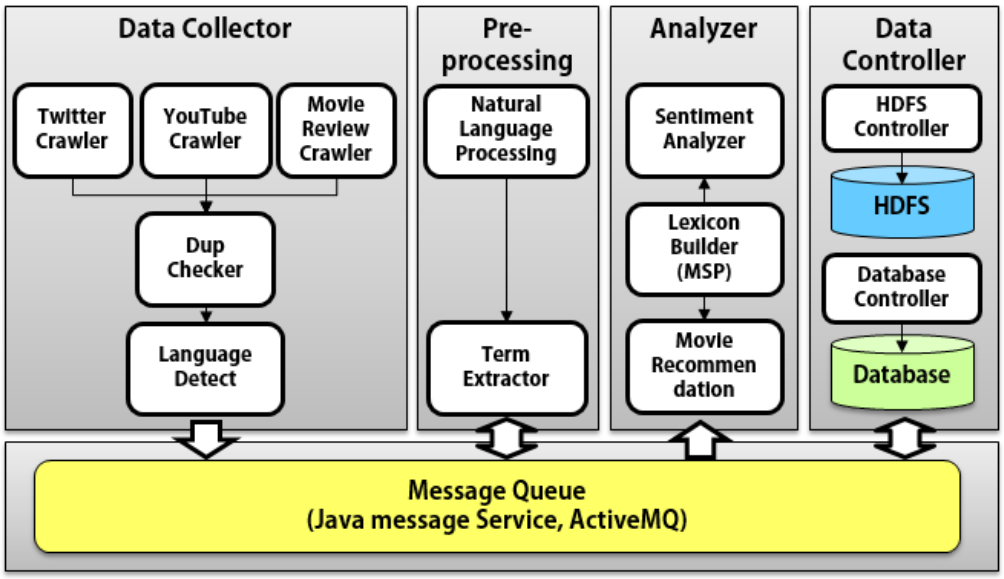

Figure 1. System architecture and flow

3.1. Data Collector

To collect tweets from Twitter, we developed a collecting tool [68]. This tool consists of a seed crawler and a search crawler as shown in Figure 2.

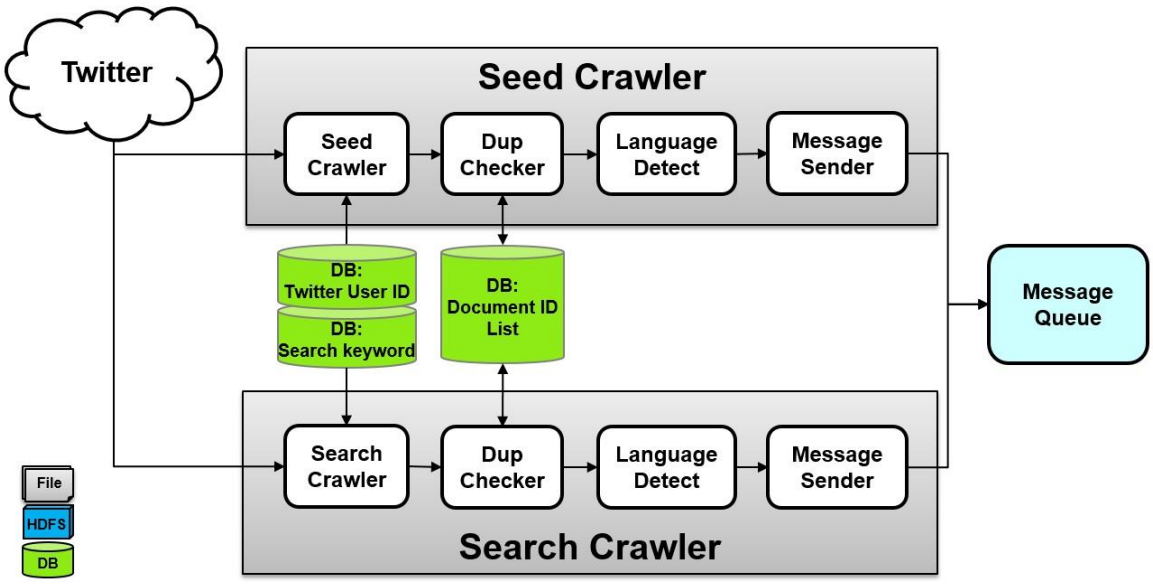

Figure 2. Twitter crawler

The seed crawler retrieves tweets based on the Twitter user's screen names as crawling seeds. The crawler automatically converts the screen names to User IDs, which consist of numeric characters. The IDs are permanently used as identifiers, while the screen names can be changeable. All the account information is stored in the database including screen names and User IDs. Twitter provides up to the most recent 3,200 tweets by each ID.

The search crawler retrieves tweets by keywords as crawling seeds. We used company names, brand names, and product names as search keywords. This crawler requests tweets using Twitter's search application program interface (API) with these keywords. Then, Twitter returns tweets when a tweet contains the keywords. Twitter provides recent tweets published in the past 7 days by each keyword. In addition, Twitter requires the Twitter development application key for both the seed crawler and the search crawler. An application key allows 180 requests per a key within 15 minutes and Twitter returns 200 tweets per a request.

To collect YouTube comments, we used a YouTube collecting tool, which was developed by Lee et al [46]. YouTube provides API to access data such as video information, user profiles, and comments written by users. This tool collects comments posted on movies, which are related to the target objects such as companies, products, politicians or movies chosen as seeds. The crawler collects the data repeatedly within scheduled time.

To collect movie reviews from IMDb, Rotten Tomatoes, and Metacritic, we developed a collecting tool using the jsoup HTML parser, which is an open-source Java library designed and developed to extract information stored in HTML-based documents by Jonathan Hedley [10]. The crawler collects movie reviews and their ratings using movie names as seeds such as "Jurassic World" and "Avengers: Age of Ultron." Each movie review site has a different rating scale. In the case of Rotten Tomatoes, writers

indicate their opinions whether “Fresh” as a positive, or “Rotten” as a negative. In the case of IMDB and the Metacritic, writers indicate their opinions on scale of 1 to 10 point(s). The bigger number has a more positive polarity. Thus, we decided 8 to 10 are positive opinions and 1 to 3 are negative opinions to calculate positivity of expressions.

To prevent data duplication, we developed a duplication checker. All the data should be checked before being sent to the next phase, because the data providers do not prevent data duplication from collecting requests [68]. In addition, we developed a language detector for filtering out the non-English data using Microsoft Language Detection Module, because we only consider data written in English [68].

3.2. Pre-processing

The first step of this phase is the natural language processing. The system analyzes the base form of the word (lemma), the part of speech (POS), and the sentences structure (sentence splitting) using the Stanford Core NLP tool [40]. Especially, the part of speech tags are fundamental information for our lexicon building method named as the morphological sentence pattern (MSP) model. We generate morphological sentence patterns using the parts of speech information, tagged by the NLP tool for extracting aspects and expressions. NLP is a time-consuming task. Therefore, the system performs this task prior to the analysis phase. In addition, the system filters out meaningless

sentences, which are not containing either a noun, an adjective, or a verb. In the linguistic approach, these are considered as meaningless sentences [10, 11]. The system also filters out and transforms stop-words and reserved words, which are defined by service providers such as hash tags (#), accounts (@) or URL formats (http://) because these words may cause inaccurate results on the further analysis.

In addition, we developed a term (keyword) extractor to extract keywords from collected data using the results of the NLP tool. The term extractor finds keywords based on specific part of speech (POS) such as noun, pronoun, adjective, verb, and adverb because these POSs may have meaningful information in a document [10, 11]. We selected the top 100 keywords that are most frequently appeared in each topic in daily. Then, the extractor filters stop-word, and then, sends the data to the message queue with all the extracted results.

3.3. Analyzer

The analyzer consists of three main parts, which are a lexicon builder, a sentiment analyzer, and a movie recommendation. First, the lexicon builder extracts aspect-based lexicon (keywords) using the morphological sentence pattern (MSP) model [10, 11]. Second, the sentiment analyzer discovers emotional states, which indicates whether a positive or a negative from collected data. Third, we proposed an extracting method of

movie genre similarity for movie recommendation. This method calculates movie genre similarity based on two main methods, which are “TDF-IDF” and “Genre Score,” using the aspect-based lexicon. In section 5 and 6, we describe each analysis method in detail.

3.4. Data Controller

To handle massive data generated from social media, we developed a data controller as shown in Figure 3. We use Apache Hadoop for a data repository and ActiveMQ, which is a messaging service for asynchronous data communication. Apache Hadoop is an open-source software for data processing and management with distributed manners [103]. This system allows distributed processing of large data sets across clusters of computers with simple programming models and it supports job scheduling and cluster resource management tools. The distributed computing techniques provide high scalable processing capabilities that reduce processing time for big data [104]. Hadoop uses Hadoop Distributed File System (HDFS) that splits files into large blocks and distributes blocks into cluster nodes. We use Map-Reduce to calculate keyword rankings. Map-Reduce is a distributed data processing model, and an execution environment that processes data in parallel within the nodes. Hadoop is widely used to analyze big data, therefore we use Hadoop as a data repository.

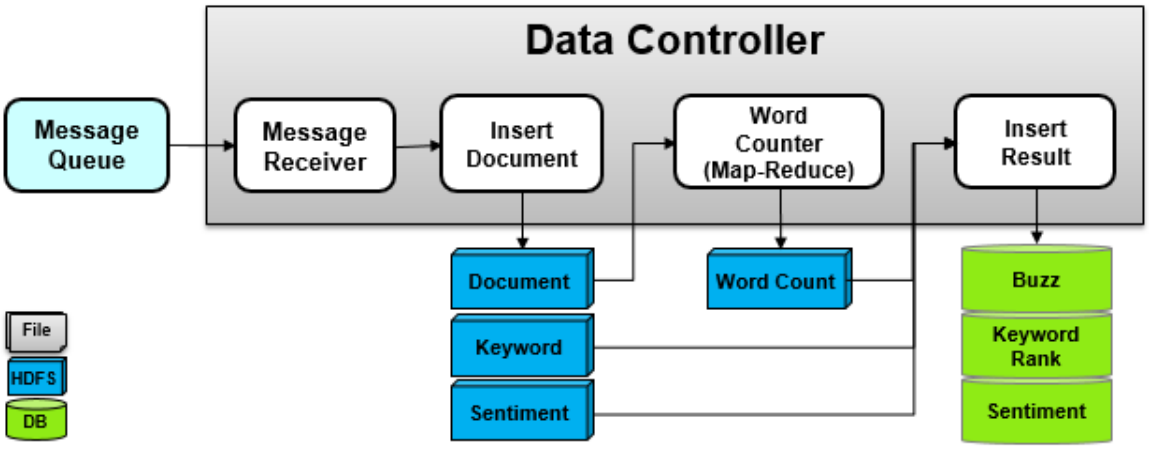

Figure 3. Data controller

A messaging service is a loosely coupled and distributed data communication service between sub-systems or modules. Message-oriented middleware supports asynchronous communication between distributed components such as senders and receivers using a message-passing. In this model, a sender sends messages to designated queues, then a receiver receives the message from the queue asynchronously.

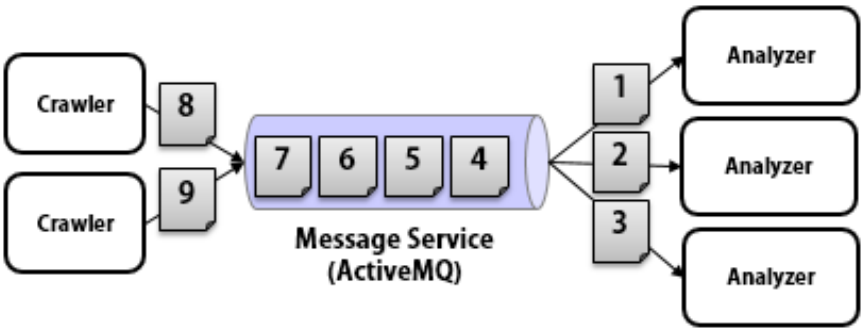

Figure 4. Message service of point-to-point model

The Java Message Service (JMS) is a vast collection of application program interface (API) to develop a message-oriented service. JMS supports two models, point-to-point and publish-subscribe. In the point-to-point model, messages are passed to an individual consumer, which maintains a queue of incoming messages. On the other hand, the publish-subscribe model is based on topics that can be subscribed to by clients. In this model, messages are sent to a topic queue, then all the subscribed clients receive the same messages. In addition, JMS supports the asynchronous mode for flexibility of communication. In this mode, sender(s) and receiver(s) are not necessarily fully-connected to communicate messages. All messages are stored in the queue until all client receive the messages. JMS is widely used to integrate various platforms and it helps to reduce bottlenecks and increase scalability [105, 106, 107]. In this research, we used the point-to-point model for multi-processing as shown in Figure 4.

3.5. Experiment Result

In this section, we examine our system with a case study. We selected 22 companies to analyze from S&P 100, a stock market index of United States stocks maintained by Standard & Poor's. The 22 companies were categorized into three groups by their business types as shown in the Table 2.

Table 2. Selected companies for experiments

Category	Company
Information Technology	Apple, Amazon, Cisco, eBay, Facebook, Google, HP, IBM, Intel, Microsoft, Oracle, Qualcomm
Retail	CVS, Costco, Home Depot, Lowe's, Walmart, Target, Walgreen
Telecommunications	AT&T, Verizon

To collect the data, we assigned 6 app keys for seed crawler. 6 app keys allows 1,080 queries per 15 minutes. Because a query includes 100 tweets, 108,000 tweets can be crawled in every 15 minute. It means that the crawler can collect 7,200 tweets per a minute. We also assigned 28 app keys for search crawler. The 28 app keys allow 5,040 queries per 15 minutes. So, 504,000 tweets can be crawled in every 15 minutes. This means that the crawler can collect about 33,600 tweets per minute. Table 3 shows examples of keywords and user accounts to collect data.

Table 3. Examples of keywords and user accounts for collecting

Category		Count	Example
Keyword	Company	55	AT&T, Csco, Cisco, Home depot,
User account	Company	409	AT&T Business (@ATTBusiness), Cisco Services (@CiscoServices), Home Depot Inc (@HomeDepot)
News		119	Wall Street Journal (@WSJ), Reuters Business (@ReutersBiz), USA today (@USATODAY)
Investment company		98	Fidelity Investments (@Fidelity) Vanguard (@Vanguard), Goldman Sachs(@GoldSachsNews)

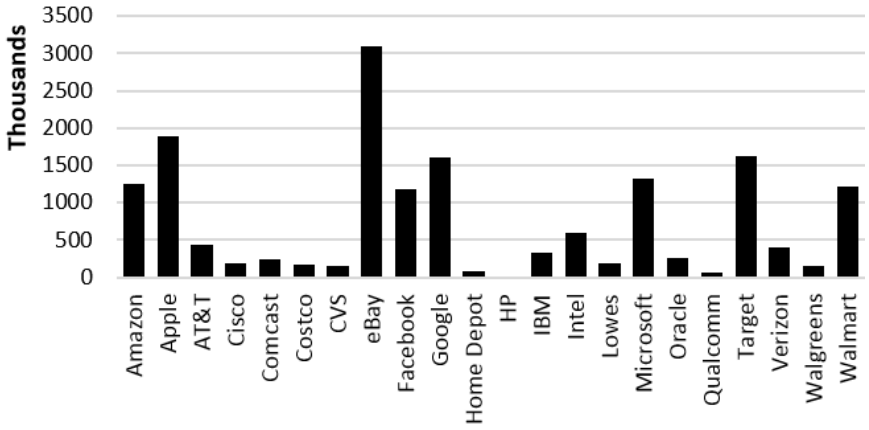

Figure 5. The number of data by the company

From Dec 19, 2014 to Jan 18, 2015, the crawler collected 16,479,483 documents. On average, that is about 532,226 documents daily. We can expect that about 2 billion documents will be collected per year through this system. Figure 5 shows the total amount of documents by the companies.

Table 4. Comparison of the number of processed documents by the crawler, analyzer, and data controller

Processing / 10 min	Crawler	Analyzer	Data Controller
test 1	2,109	599	1,967
test 2	2,035	566	2,149
test 3	2,004	455	2,161
test 4	2,120	489	2,294
test 5	2,485	448	2,171
test 6	2,464	443	2,155
test 7	2,593	451	2,180
test 8	2,563	491	2,152
test 9	2,698	483	2,260
test 10	2,670	616	2,727
average	2,374	504	2,222

Table 4 shows comparisons of the performance by 10 minutes per documents between the crawler (data collector), the analyzer, and the data controller. On average, the crawler handles about 2,374 tweets in 10 minutes. The analyzer handles about 504 tweets per 10 minutes. This is about 4.7 times slower than the crawler in the average as shown in the test 5. The data controller handles about 2221.60 tweets in 10 minutes. This is about 1.43 times slower than the crawler in the average as shown in test 1.

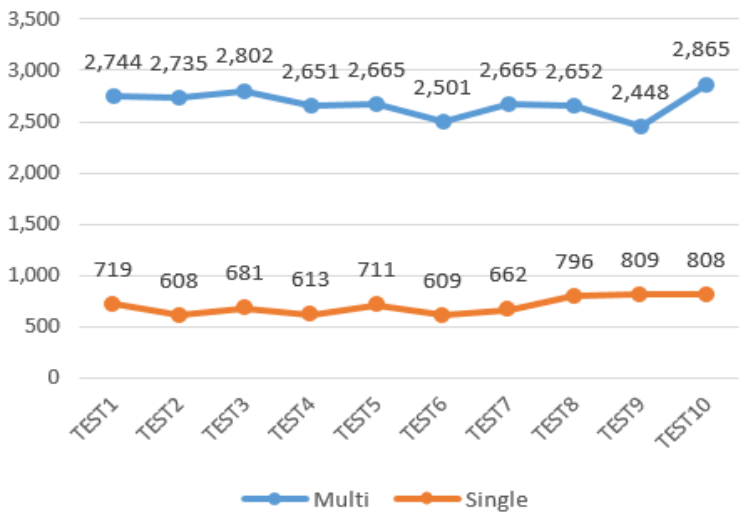

Figure 6. Comparison between multi-processing and single-processing.

According to the result, we applied multi-processing to minimize the gaps. Also, we examined how this system performs to process data in real-time. As shown in the Figure 6, results of the multi-processing is about 2,673 documents per 10 minutes on average. This means that the system can cover an average of the crawler (2,374 documents per 10 minutes in average). Thus, we can execute more processors of the analyzer and the data controller as following the results. All the stored keyword is counted by the word counter using Map Reduce technology in every 10 minutes for further analysis.

In this section, we proposed a system to discover meaningful information from social data in real time. The system is designed to support the multi-processing and the message service to improve maintainability, capability of multi-processing, and flexibility of processes controlling. Most importantly, we used Hadoop, Map Reduce, and

ActiveMQ to handle the mass amount of data. The system can collect and process about 532,226 documents daily related to the 22 target companies. The experiment results show that our approach is able to effectively process the social data in real time.

4. SOCIAL MEDIA ANALYSIS : THE COMPARATIVE ANALYSIS OF TWITTER USAGE PATTERNS

This study is interested in understanding how people send out their messages, which devices and platforms in Twitter. The fast growing technological developments provide many options for people, and we are interested in answering how they selectively use certain outlets over all other alternatives. Another area that we seek to answer in this study is the level of activity among Twitter users: whether or not people who post messages on Twitter are just posting their thoughts and emotions or actually interacting with other people. Desire to maintain social interaction has been identified as one of major motivations for media consumption [13, 31]. At the same time, literature on Twitter use reports that the interactivity among Twitter users tend to be lower than expected [32]. The following research questions have been addressed in this study:

1. What types of messages are mostly exchanged on Twitter? How engaged are people through Twitter conversation?
2. How do Twitter users post tweets? What kinds of devices and platforms do they prefer using?
3. How pervasive are tweets from business/profit-oriented sources in the Twitter world, compared to those from individual consumers? How different are these tweets in terms of devices and message types?

In addition, most research on social TV was focused on reality TV shows or live political events, and there is few work on a sporting event or game. Therefore, we also focused on one of the most popular sporting games, Super Bowl. Within the framework of previous social TV literature, this study addressed the following research questions for a better understanding of social TV viewing experiences with the Super Bowl Game.

The following research questions have been addressed in this study:

4. How relevant are tweets to a sport TV program in terms of topics?
5. What communication activities do Twitter users engage in while watching a sport TV program?

4.1. Method

Instead of using a traditional content analysis, this study used data-mining techniques to collect and comprehend massive amounts of tweets exchanged, such as the frequency and amount of traffic generated, types of tweets, devices used to post tweets, and the nature of tweets exchanged for a certain brand or commercial.

As survey, experiment, and content-analysis methods are conventional research methods in the mass communication field, the data-mining techniques are well-adopted in the computer science field and allow researchers to handle a huge amount of data and discover knowledge and information from them [69, 70, 71]. As the amount of tweets exchanged in cyber space is enormous, it is not possible to retrieve, filter, analyze, and

visualize them without automated-tools and well-defined approaches. For this reason, all tweets exchanged in the study period and user profiles were collected through an automated Twitter data collecting tool which was developed from previous studies [70, 71].

Among numerous tweets, we decided to examine the tweets about Super Bowl games and advertising, mainly because the Super Bowl game was one of the five most-watched broadcasts in 2012 (with 111.3 million viewers) and became the most watched TV program in U.S. television history in 2014, reaching 112.2 million viewers [39, 72]. With the average cost of $4 million for a 30-second spot, the Super Bowl remains a major advertising venue due to its ability to draw such a large audience and create buzz about the commercials and brands. With the popularity of Twitter, the Super Bowl becomes an important venue for generating active tweets and communicating with consumers. For example, in 2011, the Super Bowl was ranked the third highest TV event in terms of a social-TV total activity score, right after England's Royal Wedding and the MTV Video Music Awards [73]. Here, the social-TV total activity score measures social media activity related to major TV programs on Twitter, Facebook profiles, and the social applications, such as GetGlue and Viggle [74]. Particularly, Super Bowl XLVI in 2012 was marked as the first attempt at converging social media and television broadcasting that successfully drew viewers' interests. During the 2014 Super Bowl game, some marketers such as Pepsi, Samsung, and Oreo, formed a "mission control" center or a "war room" at their companies to monitor Twitter messages and to interact with their

audiences [75]. All of these suggest that the Super Bowl is a relevant venue to draw tweet messages for the analysis.

The study period was the three weeks surrounding the Super Bowl game in each year of 2012, 2013, and 2014: one week before and two weeks after the Super Bowl (Jan. 29, 2012 to Feb. 19, 2012; Jan. 27, 2013 to Feb. 17, 2013; Jan. 26, 2014 to Feb. 16, 2014). This study period was chosen to include all tweets related to the topic since marketers released their ads on social media sites like YouTube prior to the actual broadcast of the game in hopes of creating more buzz, and Twitter traffic is typically higher than average for a few weeks after the game, as the lingering impact of the advertising continues [76].

To address the research question 1, 2, and 3, we analyzed two sets of data: "General Data" and "Sample Data." "General Data" included all tweets exchanged during the study period (Jan. 29, 2012 to Feb. 19, 2012; Jan. 27, 2013 to Feb. 17, 2013; Jan. 26, 2014 to Feb. 16, 2014). Out of this data set, a "Sample Data" set consisted of Super Bowl commercial related tweets. Tweets about Super Bowl commercials were retrieved from this "general data" set by using key words, such as "Super Bowl," "Super Bowl commercials," "ads," and any company/brand name or commercial titles that were broadcasted on each of three Super Bowl games. For example, key words such as "Pepsi," "Soundcheck," "Bud Light," "Epic Night," "Jeep," "Restlessness," "Hyundai," "Sixth sense," "H&M," "David Beckham," and "NFL" were used. The unit of analysis

was every single tweet identified by the aforementioned search terms within the study period. Overall, we retrieved 73,192 tweets (35,187 in 2012, 34,350 in 2013, and 3,655 in 2014) related to the Super Bowl commercials.

Following the typology suggested in previous studies [77, 78], each tweet in the data set was classified into three categories: Singleton, Retweet, and Reply. A Singleton is classified as an undirected message, where no specific recipient is suggested. So when a user posts a tweet without referring to other users or tweets, we classified it as a Singleton. When a user sends a tweet by reposting someone else's tweet, it is called a Retweet and is marked by the prefix "RT." All tweets with RT were classified as a Retweet. When a user posts a tweet by referring to another user with an @ sign, it is considered a Reply. A Reply message is different from other categories in that it sends a tweet to a designated person. All messages with an @ sign were classified as a Reply. Thus, among these three types, a Reply is considered a higher-level engagement between users than the other two, while a Retweet is considered a lower-level message exchange between users in that a user simply reproduces a tweet written by another user without further adding his/her own messages. For that reason, we examined the percentages of a Retweet and Reply in the data "sample" to analyze the degree of the message exchanges between users. The higher percentages of Reply would indicate that message exchanges and engagement were made at a higher level among Twitter users while the higher percentages of a Retweet or Singleton indicate a lower-level exchange.

Twitter provides the name of the platform which contains specific uniform resource locator (URL) information, showing how each tweet was posted. Three graduate students analyzed all URLs used to tweet and found that about 99.8% of all tweets were generated from 600 URLs. These identified 600 URLs were used for source analysis, and their examples were listed in Table 5.

Table 5. Examples of URL sources and tweet counts

Source	URL	Tweet count
Twitter for iPhone	http://twitter.com/download/iphone	569,018
Twitter for Android	http://twitter.com/download/android	352,537
Twitter Official Web	http://twitter.com	311,690
Twitter for BlackBerry	http://blackberry.com/twitter	74,061
Twitter for iPad	http://twitter.com/#!/download/ipad	53,637

To address the second research question, each tweet was coded into two mediums for the type of medium used to post: mobile and desktop. Here, a tablet was included as mobile. Then, a mobile device is further classified into three categories since Twitter users have three options to tweet from their mobile devices: as Twitter official sources (Twitter mobile applications and Twitter official mobile web site), business sources (business and profit oriented sources), and miscellaneous (other mobile applications, 3rd party web site, or unknown sources). In the same way, the tweets posted through the desktop computer were further classified into three categories: as Twitter official sources (Twitter official desktop web site), business sources (business and profit oriented sources), and miscellaneous (other mobile applications, 3rd party web site, or unknown sources).

Table 6. Categories of tweet sources for coding

Device	Source	Example
Mobile	Twitter Official	- Twitter for iPhone, - Twitter for iPad - mobile.twitter.com
	Business	- TwitRocker2 - Tweetro+
	Miscellaneous	- Instagram
Desktop	Twitter Official	- Twitter Official Web
	Business	- Unfollowers.me
	Miscellaneous	- Facebook

Table 6 summarizes the categories used to analyze the device and platform preference in posting tweets. Here, a business source means a website domain that is owned by private companies to provide Twitter related advertising business or Twitter analysis services. These business sources are profit-oriented sites that support marketing efforts for various companies and organizations by contracts. Table 7 shows profiles of the Top 10 official and Top 10 business sources.

Table 7. Profiles of the Top 10 Official Sources

Name	URL	Type
Twitter for iPhone	http://twitter.com/download/iphone	Mobile
Twitter for Android	http://twitter.com/download/android	Mobile
Twitter Official Web	http://twitter.com	Company
Twitter for BlackBerry	http://blackberry.com/twitter	Mobile
Twitter for iPad	http://twitter.com/#!/download/ipad	Mobile
Twitter for Android	https://twitter.com/download/android	Mobile
Tweet Button	http://twitter.com/tweetbutton	Application
Mobile Web (M2)	https://mobile.twitter.com	Mobile
Twitter for Windows Phone	http://www.twitter.com	Mobile
Mobile Web (M5)	https://mobile.twitter.com	Mobile

To address the research question 4 and 5, we analyzed two sets of data included all tweets exchanged during the Super Bowl XLVII held on Sunday, Feb. 3, 2013. In this game, 108.41 million viewers tuned in to Super Bowl XLVII and the game earned an average overnight household rating of 46.3, making it the second-highest-rated Super Bowl in 27 years [38, 39]. Along with this high rating points, the 2013 Super Bowl XLVII (scored 34-31) was dynamic with several NFL records, such as the touchdown by a 190-yard kick-return (4th touchdown), a touchdown by a quarterback, and a 34-minute blackout due to a power outage. Therefore, the 2013 Super Bowl game would provide a good case material to examine multimedia experiences between television and Twitter.

Accordingly, all tweets exchanged between 6:00 p.m. and 11:59 p.m. on Feb. 3, 2013 were identified, and several keywords (e.g., Super Bowl game, touchdowns, and field goals, and names of players and coaches) were used to select sample tweets on Super Bowl game. Table 8 shows all the terms used to draw the samples. Only tweets written in English were collected and analyzed. The current open access policy to Twitter data through its own APIs allowed us to identify, collect, and analyze the data with the algorithm we developed. All tweets were grouped by minute to track the movements of the tweets during broadcasting. Then, the total number of tweets per minute was visualized with line graphs. To see how tweeting has changed over the course of the game, the number of total tweets was compared against major turning points or events (e.g., touch downs, field goals, and the half-time show) (see Figure 7).

Table 8. List of filtering keywords used to identify Super Bowl-related tweets

Categories	Keywords
Super Bowl 2013	Super Bowl, Baltimore, Ravens, San Francisco, 49ers, Beyoncé, commercial, New Orleans, Mercedes-Benz Superdome
Related People	John Harbaugh, Joe Flacco, Ray Lewis, Naquin Boldin, Dennis Pitta, Jacoby Jones, Justin Tucker, Jim Harbaugh , Frank Gore, Colin Kaepernick, Michael Crabtree, David Akers
American Football Terms	Touch down, field goal, kick, pass, rush, receive, run, tackle, sack, down, penalty, safety, halftime, kickoff, fumble, score

The communication activities made by viewers were coded into two categories: "Self-expression" and "share." As discussed earlier, previous literature identified various motivations and communication activities that TV viewers have shown. Among those, we simplified these communication activities into the two categories based on Twitter's own nature. The nature of tweeting itself contains the intention to share with others and Twitter is designed to enhance interactivities among users. However, people simply post tweets just to express themselves as Shao [22] revealed. Self-expression means that users had a strong intention to express their feelings/ideas/opinions. "Share" refers to a user's intention to inform or share them with others. If tweets have RT (retweet someone else's tweets), @ (directly delivered to a designated user), links/URLs, or state facts and information, these tweets were coded as "intention to share." The tweets without @, or RT were examined and categorized into either "self-expression" or "share," by reading each message. "Self-expression" tweets were likely to be emotional or evaluation-related. In this process, we found the tweets with a hashtag # were tricky to code. Even if a

hashtag's key function is to facilitate the search or sharing of tweet messages, we found that those with # cannot be automatically categorized into "share." We examined each tweet with a hashtag and coded it either "self-expression" or "share." To develop a code book, three authors used different tweets (not related to this research) and ran pretest three times with three different sets of data to build a clear understanding of this variable (e.g., self-expression vs. share). Once we established strong agreement with this pretest, two authors took 50 samples, which yielded a Cohen's Kappa of .97.

Previous studies used two different types of programs to examine how tweets during broadcasting were relevant to broadcast program content. Instead of using two different TV programs, this study used two different broadcast period during the Super Bowl game: Blackout period and Halftime Show. Blackout lasted about 34 minutes during the 2013 game and a Halftime Show ran 30 minutes. So, these two are compatible in terms of length of TV broadcasting time. To measure how relevant tweets were to a Halftime show, we counted any tweets mentioned with halftime, Beyoncé, her band, or her performance as "relevant" (or matched). The remaining tweets during a Halftime show was coded as "not relevant" ones. Tweets during a Blackout period were coded in a similar way. If tweets were about blackout, power outrage, light, or darkness, they were coded to "relevant" (matched) tweets. Others were coded as "not relevant" tweets. We took a similar step to establish a high agreement on this variable through a pretest. Then, 50 sample tweets were coded by two authors, which yielded a Cohen's Kappa of .98.

Table 9 shows a list of keywords used to identify the relevance of each tweet to Halftime Show or to Power Outage.

Table 9. List of keywords used to filter tweets related to non-scoring events

Categories	Keywords
Halftime Show	Halftime show, Beyonce, Kingbey, Destiny's child
Power Outage	Power went out, Blackout, Electric bill, Gotham, Light out, Bane, Undertaker

4.2. Experiment Result

A total of 1,413,524 tweets in 2012, 2,079,902 tweets in 2013 and 1,852,181 in 2014 were retrieved during the study period. Out of these ("General Data" of 5,345,607 tweets), a total of 73,192 Super Bowl related tweets (called "Sample Data") were analyzed (i.e., a total of 35,187 tweets in 2012, 34,350 in 2013, and 3,655 in 2014). Over these three years, the total number of Super Bowl related tweets was the lowest in 2014, and the portion of Super Bowl commercial related tweets out of total tweets significantly decreased to 0.2% in 2014, compared to 2.5% in 2012 and 1.7% in 2013. It seems that there were fewer messages and conversations on Twitter about Super Bowl commercials in 2014, compared to the two previous years (See Table 10).

Table 10. Data "General Data" and "Sample Data" by year (2012~2014)

Years / Type	2012	2013	2014	Total
The number of "General Data" (All tweets)	1,413,524	2,079,902	1,852,181	5,345,607
The number of "Sample Data" (Super Bowl commercial related tweets)	35,187 (2.5%)	34,350 (1.7%)	3,655 (0.2%)	73,192

The first research question asked what types of messages were mostly exchanged on Twitter. Over the past three years from 2012 to 2014, a Singleton is the most popular message type (accounting for 37-42%), but all three types show relatively similar

portions. The year of 2014 had a lower portion of Reply (accounting for 17.8%), compared to the two previous years (32.2% in 2012 and 28.3% in 2013) (See Table 11). However, when we examined the tweets related to Super Bowl commercials, we see a much wider differences among the three different message types. Singleton messages accounted for a larger portion (about 60%-72%), followed by Retweet (17%- 35%), and Reply (6%-13%).

Table 11. Type of messages by year (2012~2014) in "General Data"

Years / Type	2012	2013	2014
Singleton	534,990 (37.8%)	763,470 (36.7%)	778,726 (42.0%)
Retweet (RT)	423,138 (29.9%)	727,717 (35.0%)	743,926 (40.2%)
Reply (@)	455,396 (32.2%)	588,715 (28.3%)	329,529 (17.8%)
Total	1,413,524	2,079,902	1,852,181

Even if each year showed a different portion of each message type, the overall pattern was consistent. The percentages of Retweet consistently increased from 2012 to 2014, while the portion of Reply tweets has significantly decreased. In 2012 and 2013, Reply tweets accounted for 12.4% and 10.8% of all tweets, respectively (See Table 12). However, it was drastically decreased to 5.5% in 2014. These differences were statistically significant ($\chi 2=766.01$, $df = 4$, $p < .0001$). This finding suggests that people used three different types of messages almost evenly as they tweeted general messages, but when they tweeted about Super Bowl commercials, they mostly used a Singleton type by posting undirected messages without much interactions with other users.

Table 12. Type of messages by year (2012~2014) in "Sample Data"

Type \ Years	2012	2013	2014
Singleton	24,971 (71.0%)	22,733 (66.2%)	2,293 (60.3%)
Retweet (RT)	5,850 (16.6%)	7,899 (23.0%)	1,253 (34.3%)
Reply (@)	4,366 (12.4%)	3,718 (10.8%)	200 (5.5%)
Total	35,187	34,350	3,655

($\chi 2$=766.01, df = 4, p <.0001)

The second research question intended to find tweeting methods: how users post tweets and the device/platform that they use. As Table 13 shows, Twitter users preferred mobile devices (1,218,594 tweets, 65.9%) to desktop computers (629,160 tweets, 34.1%) as they tweeted general messages. Even if there are more sources for Twitter posting on desktop computers (469 sources, 78.2%) than in mobile devices (131 sources, 21.8%), two thirds of all tweets were generated via mobile devices. This means that the majority of people prefer the mobile device to post general tweets. When we examined the Super Bowl related tweets (i.e., "Sample Data"), however, we found a different pattern. A desktop was used for posting 67.8% of Super Bowl related tweets (2,399 tweets) while a mobile device was for posting 32.2% of Super Bowl tweets (1,139 tweets). This was an unexpected finding, but as we examined this data with the sources, this unexpected finding was understood. It might be due to the fact that 85% of tweets posted through mobile devices came from official sources (966 tweets out of 1,139 tweets), while none of tweets came from business sources. On the other hand, only 46% of tweets posted through desktop devices came from official sources (1,095 tweets out of 2,399 tweets)

and the remaining portion of tweets through desktop devices (54%, 1,304 tweets out of 2,399) was from business profit-oriented sources or miscellaneous sources. Also, all Super Bowl related tweets generated from business sources were posted through desktop devices (all 578 tweets). In other words, all business profit-oriented sources pushed Super Bowl related tweets through desktop devices, which might have contributed to a higher number of desktop device data. This complicated data was further analyzed and discussed in a later section.

Table 13. The number of tweets generated through devices and sources (2014)

Source / Type	Mobile				Desktop				Total
	Official	**Business**	**Misc.**	**Total**	**Official**	**Business**	**Misc.**	**Total**	
The number of sources	14	2	115	131 (21.8%)	18	88	363	469 (78.2 %)	600
The number of "General Data" (All tweets)	1,157,016	133	61,445	1,218,594 (65.9%)	343,999	120,665	164,607	629,271 (34.1%)	1,847,865
The number of "Sample Data" (Super Bowl commercial related tweets)	966	0	173	1,139 (32.2%)	1,095	578	726	2,399 (67.8%)	3,538

Once a device was identified, we examined the sources that generated tweets. The sources were categorized into the three groups mentioned earlier in the method section (see Table 6): Twitter official sources, business sources and miscellaneous sources. Among all identified 600 URLs as twitter posting sources, 32 URLs were Twitter official

sources, 90 URLs were business sources, and 478 URLs were miscellaneous sources. Even if Twitter official sources were smaller, most tweets were generated from Twitter official sources (1,501,015 out of 1,847,865 general tweet messages and 2,061 out of 3,538 Super Bowl related tweets). This means that even though there are many applications or web pages where Twitter functionalities are integrated, the majority of people prefer to post tweets through the Twitter official sources.

Also, among all identified 90 business sources, 88 URLs and 99.9% of general tweets (120,665 tweet out of 120,798 tweets) came through desktop business sources, and only 133 tweets (0.1%) were generated through mobile business sources (see Table 13). This means that most of business and marketing promotional tweets were generated through desktop devices.

Table 14. Type of tweets generated through Twitter official and business sources from "Sample Data"

Source / Type	Official Sources	Business Sources
Singleton	780 (37.8%)	513 (88.8%)
Reply	182 (8.8%)	4 (0.7%)
Retweet	1,099 (53.3%)	61 (10.5%)
The number of "Sample Data" (Super Bowl commercial related tweets)	2,061	578

To examine whether the nature and message types of Super Bowl related tweets were different between official Twitter sources and business sources, the three message types were further examined by the sources that each message was produced from. Out

of all 2,061 tweets generated through Twitter official sources, more than half of these tweets were Retweet (53.3%) and Reply (8.8%), while 37.8% of them were Singleton. On the other hand, among 578 tweets generated through business sources, 88.8% of them were Singleton only with small portions of Retweet (10.5%) and Reply (0.7%) (See Table 14). This means that users who used official Twitter sources tend to have more interactions with other users by sending out more Retweets or Reply than Singleton, while users who used business sources tend to generate one-directional (e.g., Singleton) tweets. Table 14 summarizes the message type of Super Bowl related tweets generated through Twitter official sources and business sources.

Table 15. The number of tweets and users by Top 10 official and Top 10 business sources from "Sample Data"

Source / Type	Top10 Official Sources			Top10 Business Sources		
	Tweets	**Users**	**Tweets per user**	**Tweets**	**Users**	**Tweets per user**
Singleton	776	689	**1.13**	373	130	**2.87**
Reply	181	173	1.05	4	4	1.00
Retweet	1,073	1,026	1.05	33	30	1.10
The number of "Sample Data" (Super Bowl commercial related tweets)	2,030	1,888	1.08	410	164	2.50

To find out whether there are differences in the amount of tweets generated by Twitter official sources and by business sources, we examined how frequently a single user generated tweets from each source. In this analysis, we only chose top 10 sources from each section: top 10 official sources and top 10 business sources listed in Table 7.

As Table 15 shows, 689 users generated 776 Singleton messages through top 10 official sources (1.13 messages per user) while 130 users generated 373 Singleton messages (2.87 messages per user) through top 10 business sources. This indicates that users who used business sources generated tweets more frequently than users who used Twitter official sources. This suggests that business sources were more active in generating Singleton tweets than in generating interacting (Retweet or Reply) tweets.

Table 16. The number of identical tweets of the Top 10 brand names by Twitter official and business sources from "Sample Data"

Source	Tweet Count	Identical tweets
Official Sources	1,338	150 (11.2 %)
Business Sources	345	238 (69%)

In this analysis, we were also interested in examining whether there were any duplicated messages. Here, we found that business sources generated more identical tweets on top 10 brand names than Twitter official sources. Sixty-nine percentage of all tweets from business sources were the same duplicated messages while only 11.2% of tweets from Twitter official sources were the same (see Table 16). It implies that those brand names mentioned through business sources were intentionally pushed by marketers to create social buzz on Twitter. Table 17 shows the examples of identical tweets generated through business sources. Two identical tweet messages, "Chrysler stands behind 'America's Import' Super Bowl ad featuring Bob Dylan," show two different

URLs which were embedded in tweet messages, but they actually came from the same source, dlvr.it.

Table 17. Examples of identical tweets from business sources from "Sample Data"

Tweet	Source
Chrysler stands behind 'America's Import' Super Bowl ad featuring Bob Dylan http://t.co/a43obiRA6g	dlvr.it
Chrysler stands behind 'America's Import' Super Bowl ad featuring Bob Dylan http://t.co/kr9NoyX0EA	dlvr.it
Learn more about the stories behind @Microsoft's #SuperBowl ad & get inspired for your own project! http://t.co/37wgTzIb79 #technologyrocks	Sprinklr
Learn more about the stories behind @Microsoft's #SuperBowl ad & get inspired for your own project! http://t.co/jANyKD5w1F #technologyrocks	Sprinklr
Learn more about the stories behind @Microsoft's #SuperBowl ad & get inspired for your own project! http://t.co/KReMHazRNE #technologyrocks	Sprinklr
Our own @mpcmi caught up with Seattle musician @SangoBeats to talk Super Bowl, great sports crowds & more: http://t.co/EU63FxrMd8	TweetDeck
Our own @mpcmi caught up with Seattle musician @SangoBeats to talk Super Bowl, great sports crowds & more: http://t.co/Np3vgTY9Mz	TweetDeck
Our own @mpcmi caught up with Seattle musician @SangoBeats to talk Super Bowl, great sports crowds & more: http://t.co/Or4kVFs3Cf	TweetDeck
Bruno Mars Pepsi Super Bowl XLVIII Halftime Show Announcement http://t.co/kQSEfk3jzl	SocialOomph
Bruno Mars Pepsi Super Bowl XLVIII Halftime Show Announcement http://t.co/KUstTjwpKA	SocialOomph

To address the research question 4 and 5, we analyzed a total of 5,453 messages during this study period. Figure 7 shows the changes of the tweets that viewers posted during the game. Plotting the volume of tweet activities over the course of the Game shows notable spikes in activity at significant moments, such as touch-downs, safety, and Blackout period. A total of 568 tweets were analyzed for a Halftime period (from 8:00 p.m. to 8:29 p.m.) and a total of 1,119 tweets were analyzed for Blackout period (from 8:39 p.m. to 9:11 pm).

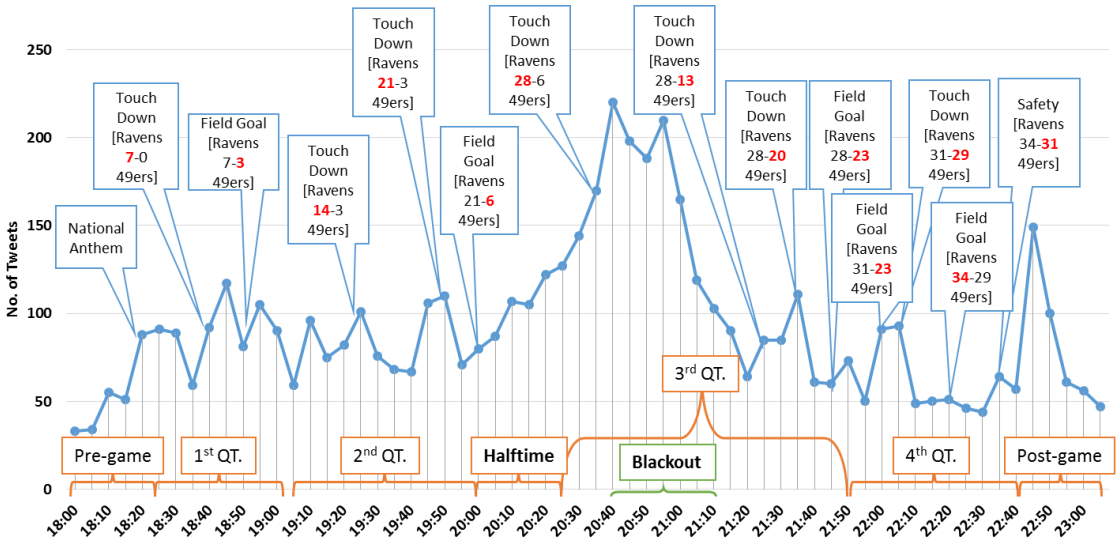

Figure 7. Tweets posted during Super Bowl 2013 as it happened per 5 min.

The fourth research question asked "How relevant are tweets to a sport TV program in terms of topics?" Table 18 shows a total number of tweets communicated during each event, and how each tweet was relevant to Halftime show or to Power Outage. There were more tweets posted during the Power Outage than during a Halftime Show. Out of 568 tweets identified during the Halftime Show, the contents of 320 tweets

(56.3%) were relevant or related to a Halftime Show while 248 tweets (43.7%) were not related to the Show. On the other hand, a larger number of tweets during Power Outage (714 tweets out of 1,119, 63.8%) was about Blackout while 405 tweets (36.2%) were not related to Blackout period. This indicates that people communicated more relevant messages about Power Outage than about Halftime Show. This difference was statistically significant ($x2 = 8.859$, $df = 1$, $p < .005$).

Table 18. Relevance of tweets for each event period

Period	Relevant	Irrelevant	Total
Halftime Show	320 (56.3 %)	248 (43.7 %)	568 (100 %)
Power Outage	714 (63.8 %)	405 (36.2 %)	1,119 (100 %)

($x2 = 8.859$, $df = 1$, $p < .003$)

The fifth research question asked "What communication activities do Twitter users engage in while watching a sport TV program?" Table 19 shows a total number of tweets communicated during each event, and how each tweet was relevant to 'Self-expression' and 'Share.' There were more 'Share' tweets posted during both event relevant tweets and irrelevant tweets. However, there were more 'Self-expression' tweets posted in the event relevant (Halftime Show-42.5% and Power Outage-33.2%) than the irrelevant (Halftime Show-27.8% and Power Outage-30.9%). Particularly, there were more 'Share' tweets (72.2 %) than 'Self-expression' tweets (27.8 %) posted in irrelevant tweets during a Halftime Show. This indicates that people tend to share tweets in

irrelevant to the events. This indicates that people preferred to share messages in irrelevant to the events. This difference was statistically significant ($x2 = 17.519$, $df = 3$, $p < .001$).

Table 19. Relevance of tweets for each event period by self-expression and share

Period	Relevant		Irrelevant	
	Self-expression	**Share**	**Self- expression**	**Share**
Halftime Show	136 (42.5 %)	184 (57.5 %)	69 (27.8 %)	179 (72.2 %)
Power Outage	237 (33.2 %)	477 (66.8 %)	125 (30.9 %)	280 (69.1)

($x2 = 17.519$, $df = 3$, $p < .001$)

To comprehend the nature of shared messages in irrelevant to events, we further analyze ‘Irrelevant-Share’ tweets posted during a Halftime show. Out of 179 tweets identified during the Halftime Show, the contents of 88 tweets (49.2%) were relevant or related to the advertisement. This indicates that the most of ‘share’ tweets are in relevant to the advertisements. Table 20 shows examples of the advertisement messages.

Table 20. Examples of "Share" tweets posted in a halftime show related to advertisements

Tweets
#Lincoln #SuperBowlad warning. RT @tbellinger: Better be careful flaunting the 45mpg #hyundai #fusion
Tell us which @twitter hashtags you are following right now! #BrandBowl #admeter #SuperBowl47
Did you spot Coach @ShaneDaniels009 in this year"s @Pepsi NEXT #Superbowl commercial?? http://t.co/LOS604Kd [Luchador&window jumper!]
RT @hilliboss: I"m glad Jeep is a major sponsor of the Super bowl! #GoJeep #GoAmerica
RT @LincolnMotorCo: Thanks! RT @autotrader_com: Well done @LincolnMotorCo! #BrandBowl #SB47
Old Spice Dances With Wolves in Super Bowl Ad Airing Only in Juneau, Alaska http://t.co/Ui2rKDXQ

4.3. Analysis

This study aimed to address the overall Twitter usage patterns by examining message types, devices and platforms used. Instead of relying on the audience's response (e.g., survey or experiment) or traditional content analysis, this study used a data-mining approach and software that are widely used in the computer science field to handle the massive amount of data. A total of 5,355,607 tweets ("General Data") and 73,192 Super Bowl commercial related tweets ("Sample Data") were analyzed.

Several findings of this study deserve further discussions. First, over the past three years from 2012 to 2014, a year of 2013 recorded the highest number of tweets, followed by 2014. This can be explained by the nature of the game as the 2013 Super Bowl was considered one of the most exciting games, yielding a final score of 34-31 with 7 touchdowns, 6 field goals, and one safety. The 2013 Super Bowl game kept the audience's attention to the last minute, making the time between 10:30 pm and 10:45 pm the most watched part of the game [79]. In addition, the game yielded several NFL records, such as the touchdown by a 190-yard kick-return (4th touchdown), a touchdown by a quarterback, and a 34-minute blackout due to a power outage, marked the dynamic nature of the game. Such a dynamic game might be a reason for the high number of tweets exchanged in 2013.

On the other hand, the 2014 Super Bowl game was recorded as the most watched Super Bowl and the most watched program in U.S. television history, reaching 112.2

million viewers [39, 72], but did not generate the highest number of tweets. This implies that the dynamic nature of the game might be a more important factor in predicting or understanding Twitter usage. This interpretation can be further supported by the finding on the Super Bowl commercial tweets. In 2014, the portion of Super Bowl commercial tweets significantly reduced to 0.2%, compared to 1.7% in 2013 and 2.5% in 2012. Even if the 2014 Super Bowl game recorded the highest viewership, the game itself was not exciting, yielding a final score of 43-8 with a big lead by the Seahawks. Such one-sided game might make people engage in other conversations or other activities rather than paying attention to a broadcasting game and commercials.

Second, the study yielded that the most popular message type for Super Bowl commercial related tweets was the Singleton. This finding implies that Twitter users are not "actively interactive" as marketers would hope to see. This confirms the previous studies that showed lack of interactivity among Twitter users even if one of the unique characteristics of Twitter as a social media platform is the interactivity [20, 22]. This has important implications for Twitter and other social media networking sites since advertisers are more likely to pay for user engagement rather than for user impressions. Marketers need to develop or add some features on social networking sites to encourage more engagement and interactions among users.

Third, the study found that individual users prefer mobile devices to desktops when they tweet, and prefer official web pages or the mobile applications provided by

Twitter even though there are many diverse applications or web pages available. This can be interpreted as people feel using Twitter's official sources safer and secure than third-party sources as well as easier to use. This finding suggests that mobile apps are good places to run any promotional messages to reach further target audiences, and Twitter's official sites or apps would be more efficient venues for marketers.

Fourth, we found that tweets generated through business sources were different from those through official sources in terms of message type, devices, and the nature. The tweets from business sources used more desktop devices, took a Singleton type, and provided more duplicated messages (69%) while tweets from official sources used more mobile devices and took a Retweet type with fewer duplicated (11.2%) messages. This was an interesting finding in understanding how Twitter is used differently between individual users and business profit-oriented users.

Fifth, we found that about half of real-time tweets are related to a sport TV program and the half of irreverent tweets are related to advertisements. As Schirra, Sun, and Bentley [34] indicated in their qualitative research on live-tweeting TV programs, live-tweeting is a complex social process with various conventions and practices. Establishing a better understanding of such practices would be an important task in the area of social TV. Our work presented here could provide a stepping stone for social TV studies on a sport game or event. We will describe an overall finding of this study in section 7.

5. SENTIMENT ANALYSIS USING MORPHOLOGICAL SENTENCE PATTERN MODEL

5.1. A Lexicon Building Method using MSP Model

People shares their emotional state and opinion with unpleasant or dissatisfied experiences about a brand, product, or person using social media. This information is a useful to understand trends, issues, individuals, human behavior, and even identifying influential people [6, 7, 8]. Sentiment analysis helps to observe and summarize the information from user-generated textual data. The aspect-based sentiment analysis is one of a lexicon-based approach. This approach performs more in-depth sentiment analysis than traditional lexicon-based approaches. As we mentioned in section 2.2, the lexicon building is one of the biggest challenge [47, 48, 49]. In this section, we propose a method for building aspect-based lexicons. This model generates morphological sentence patterns from movie reviews, YouTube comments, and tweets to extract aspects and expressions. The main purpose of this method is to minimize human-coding efforts to build aspect-based lexicons.

5.1.1. Method

Figure 8 shows the system architecture and flow of this method. The system consists of three main parts which are data collecting, preprocessing, and aspect-based

lexicon building. In the data collecting phase, the crawler collects tweets from Twitter, comments from YouTube, and movie reviews from IMDB, Rotten Tomatoes, and Metacritic (See section 3.1). In the preprocessing phase, the system filters out and refines collected data considering characteristics of social media (See section 3.2), and the system analyzes the sentence parsing (sentence splitting) and the part-of-speech (POS) tagging using the Stanford Core NLP. Then, the system extracts aspect and expression candidates. In the aspect-based lexicon building phase, the morph pattern recognizer extracts morphological sentence patterns as known as POS patterns using the candidates. Finally, the system extracts aspects and expressions using the patterns. In the rest of this section, we describe the aspect-based lexicon building method in detail.

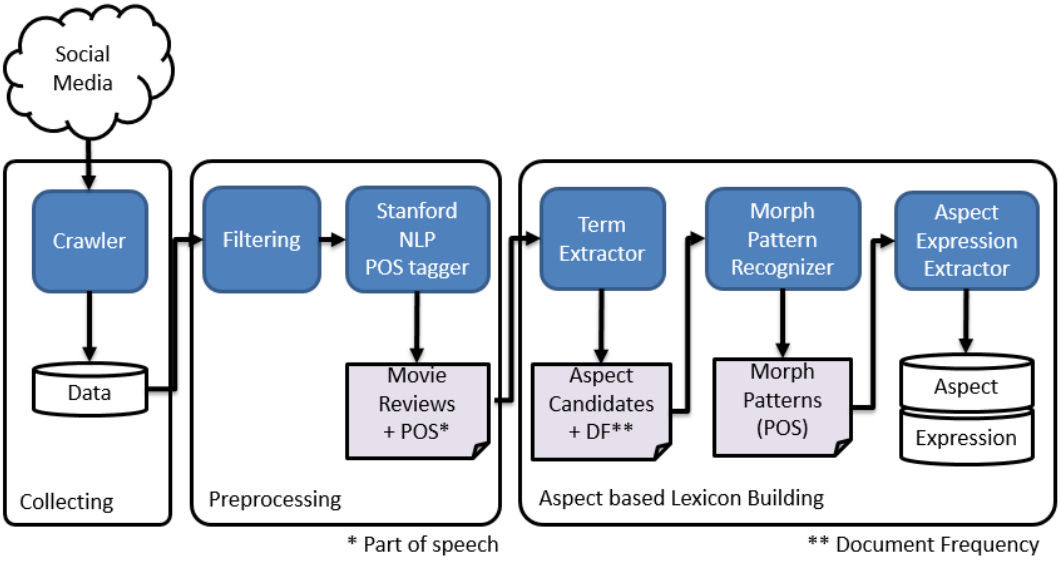

Figure 8. System architecture and flow

Once the term extractor finds aspect and expression candidates, the morphological pattern recognizer generates morphological sentence patterns in consideration of which part of speeches (POS) are surrounding the aspect and expression candidates. As shown in the Figure 9, the “character” and “sequel” are aspect candidates, which are extracted from the term extractor. The system extracts the patterns from each sentence when a sentence contains the words. Then, the system matches the patterns to extract aspects and expressions.

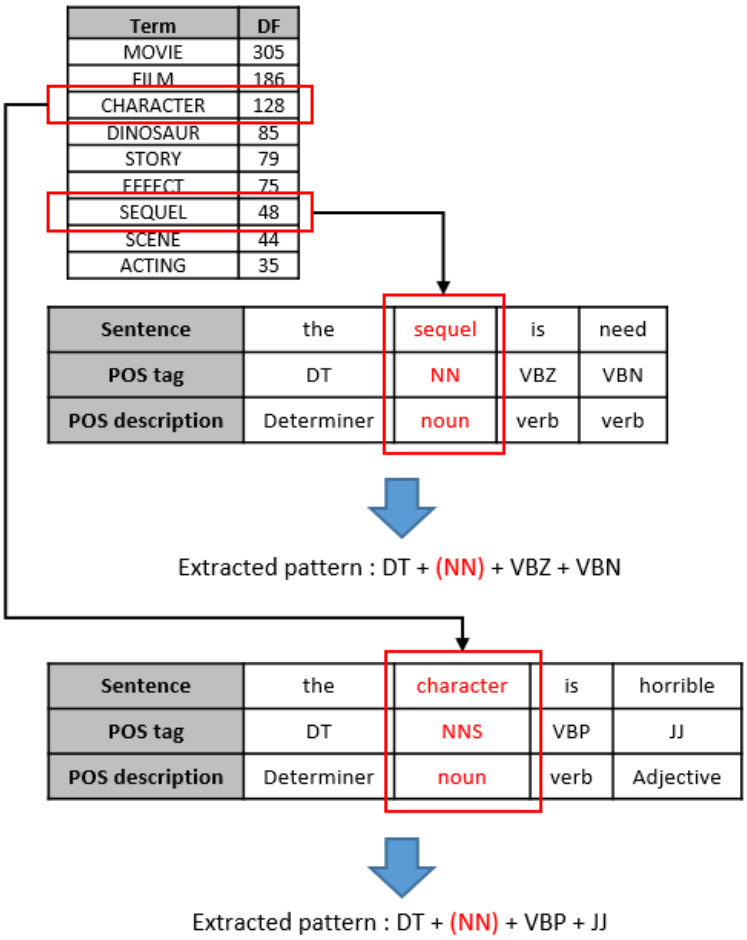

Figure 9. Example of extracting morphological sentence patterns

For a diversity of pattern matching, we applied the N-grams model, which is widely used for text-based analysis. N-gram is defined as a contiguous sequence of *n* items from a given sequence of text or speech [81]. This approach gives more possibility for pattern matching. It allows to increase the recall (coverage) for extracting aspects and expressions. Table 21 shows examples of this approach. At the first row in the table, the pattern consists of 6 length sequence of POSs (PRP$ + JJR + IN + JJ + NNP + CD + , + CC). The aspect-expression extractor matches all possible patterns in cutting each POS from the edge. In this case, we considered that the longest pattern has a higher priority to avoid duplication and this strategy helps to reduce a computation time compared to all possible pattern matching.

Table 21. Examples of a diversity of pattern matching using N-gram model

Length	**Prefix**			**Aspect**	**Postfix**		
	Pi-3	***Pi-2***	***Pi-1***	***Wi***	***Pi+1***	***Pi+2***	***Pi+3***
6	PRP$	JJR	IN	JJ NNP	CD	,	CC
5	PRP$	JJR	IN	JJ NNP	CD	,	
5		JJR	IN	JJ NNP	CD	,	CC
4		JJR	IN	JJ NNP	CD	,	
3		JJR	IN	JJ NNP	CD		
3			IN	JJ NNP	CD	,	
2			IN	JJ NNP	CD		

* P-*n*: Sequence of POSs

Once the morphological pattern recognition, the aspect-expression extractor retrieves aspects and expressions when the patterns are matched. Table 22 shows examples of extracted aspects and expressions. In the next section, we will provide experiment results.

Table 22. Examples of extracted aspect and expression candidates

Rank	Aspect	Count	Expression	Count
1	MOVIE	882	GOOD	282
2	FILM	678	BETTER	155
3	DINOSAUR	498	FUN	149
4	CHARACTER	403	ENOUGH	136
5	JURASSIC PARK	367	WELL	104
6	ORIGINAL	362	BLOCKBUSTER	99
7	JURASSIC WORLD	222	STUPID	94
8	ACTION	174	BEST	93
9	STORY	173	BIG	81
10	SEQUEL	168	BAD	80

5.1.2. Experiment Result

For the experiments, we used each 1,000 sentences collected from the movie review sites, YouTube and Twitter related to a movie, 'Jurassic World' as a seed. Table 23 shows all used data for the experiments. Even though, we used same number of sentences, the sentences per a document are totally different between sources. In the case of a movie reviews, sentences per a document is about 4 times higher (4.2 sentences in a document) than YouTube comments (1.1 sentences in a document) and 3 times higher than tweets (1.4 sentences in a document). In addition, the average number of POS in a movie review is about 2 times greater (21.1 sentences in a document) than YouTube comments (13.1) and tweets (11.3). This result shows the movie reviews consist of more and longer sentences than the others.

Table 23. Data Sample for Experiments

Sources / Category	Movie Review	YouTube Comment	Twitter Tweet
No. of Used Sentences	1,000	1,000	1,000
No. of Documents	238	911	715
Sentences / Document	4.2	1.1	1.4
Avg. of POS	21.1	13.3	11.3
No. of aspect*	230	283	96
No. of Expression*	250	341	154

*** Human-coded answer-set**

To evaluate how our method performs based on the F-measure, we built human-coded answer-sets for each site. We found 230 aspects and 250 expressions from movie

reviews, 283 aspects and 341 expressions from YouTube comments, and 96 aspects and 154 expressions from tweets. From all identified aspects and experiments, YouTube answer-set has the highest numbers of aspects and expressions. It implies that the people express opinions with more diverse expressions on YouTube than the others.

Using these data, the pattern recognizer generated 59,111 aspect patterns and 21,101 expression patterns from Movie Reviews, and 71,178 aspect patterns and 49,904 expression patterns from YouTube comments, and 47,606 aspect patterns and 36,002 expression patterns from Twitter tweets. However, we generated patterns using the N-gram model for a diversity of matching, these patterns caused too much computational time on the matching process (See Figure 16 and 17). Thus, we examined which lengths patterns can extract most numbers of correct aspects and expressions to find optimized patterns for reducing the computational time. Through our experiments, we used the average number of correct aspects and expressions as a threshold.

As shown in the Figure 10 and 11, we selected 3 to 7 lengths patterns for aspects and 2 to 6 lengths patterns for expressions from movie reviews based on the thresholds of aspects (18.2) and expressions (16.5). From all identified result, 3 to 7 lengths patterns could extract 89.4% (554 out of 620) of correct aspects and 2 to 6 lengths patterns could extract 85.2% (575 out of 675) of correct expressions from movie reviews.

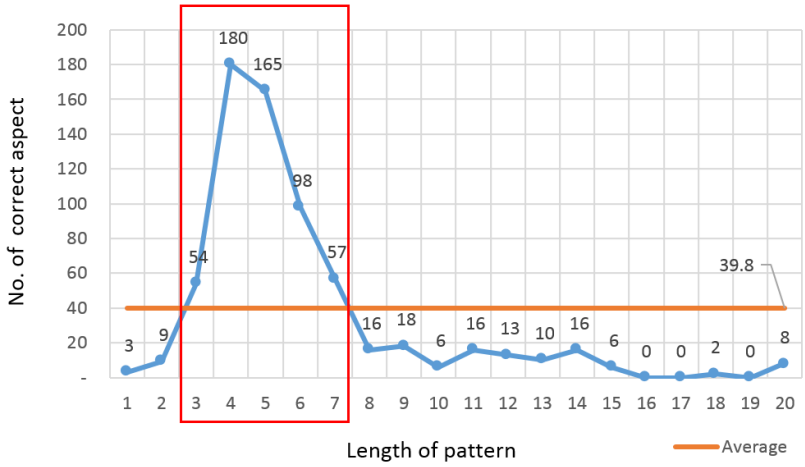

Figure 10. The numbers of correct aspects by pattern lengths for movie review

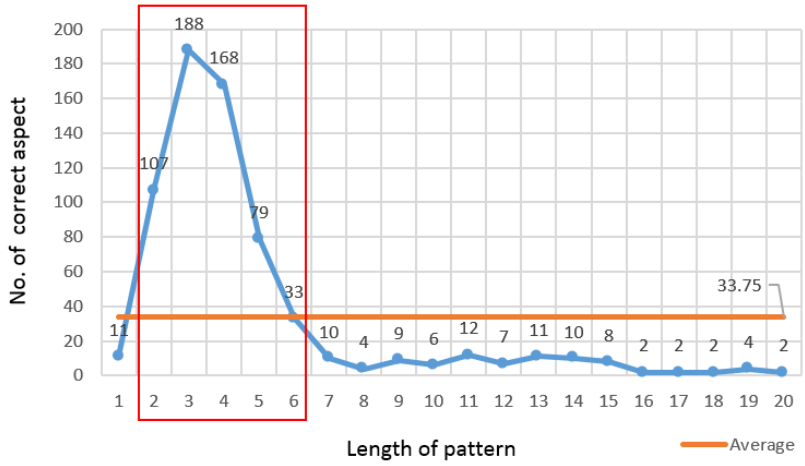

Figure 11. The numbers of correct expressions by pattern lengths for movie review

For extracting aspects and expressions from YouTube comments, we selected 1 to 5 lengths patterns for aspects and 1 to 4 lengths patterns for expressions based on the thresholds of aspects (17.9) and expressions (9.5) as shown in the Figure 12 and 13. From all identified result, 1 to 5 lengths patterns could extract 96.7% (347 out of 358) of

correct aspects, and 1 to 4 lengths patterns could extract 91.5% (173 out of 189) of correct expressions from YouTube comments.

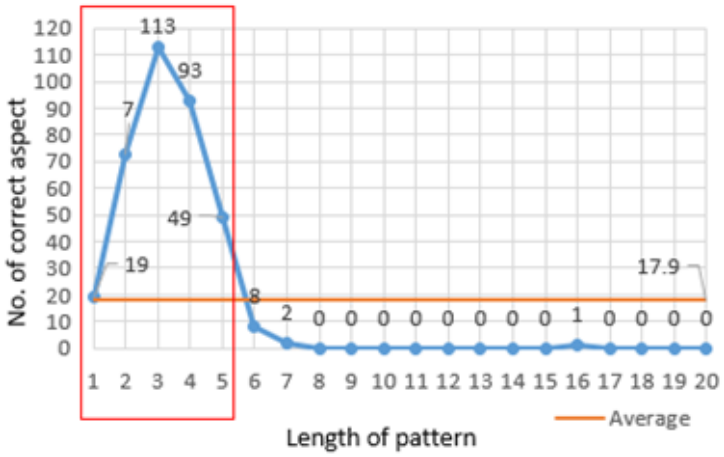

Figure 12. The numbers of correct aspects by pattern lengths for YouTube

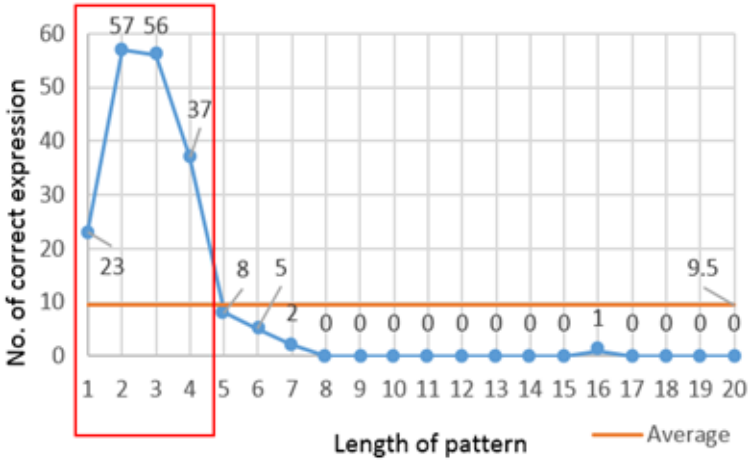

Figure 13. The numbers of correct expressions by pattern Lengths for YouTube

For extracting aspects and expressions from Tweets, we selected 2 to 6 lengths patterns for aspects and 2 to 5 lengths patterns for expressions based on the thresholds of aspects (18.2) and expressions (16.5) as shown in the Figure 14 and 15. From all

identified result, 2 to 6 lengths patterns could extract 89.5% (325 out of 363) of correct aspects and 2 to 5 lengths patterns could extract 80.84% (173 out of 214) of correct expressions from tweets.

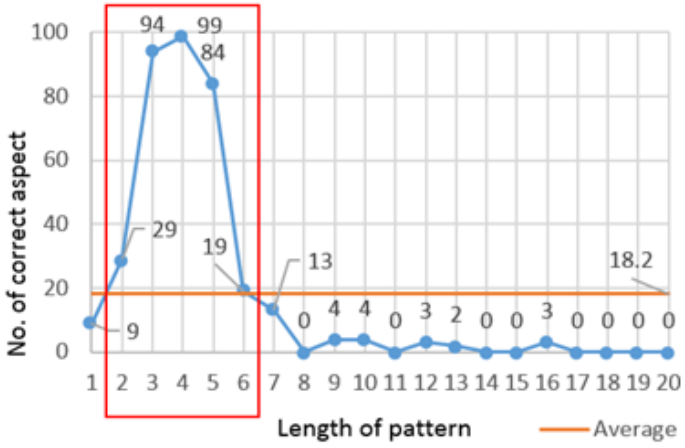

Figure 14. The Numbers of Correct Aspects by Pattern Lengths for Twitter

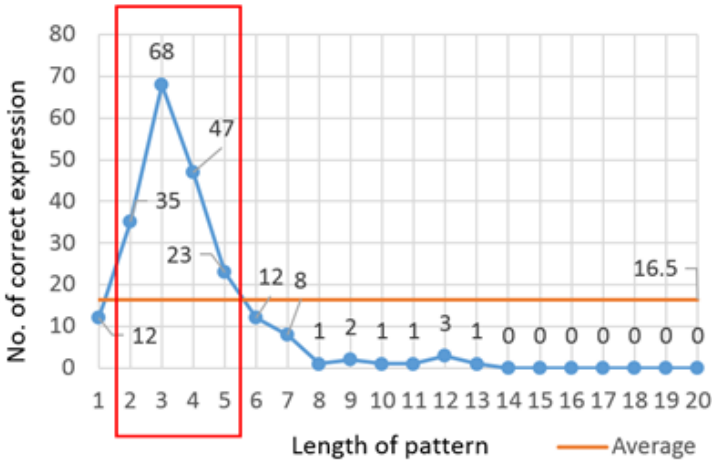

Figure 15. The Numbers of Correct Expressions by Pattern Lengths for Twitter

We named these selected patterns as “Selected Pattern” to experiment our proposed methods. In addition, we defined a matching method named as ‘LF’, which

mean the longest pattern first matching method. This method aims to avoid duplications of aspects and expressions because the patterns are generated from original patterns.

To examine the performance of aspect extraction, we compared all proposed methods, which are “Noun Phrase,” “Aspect Pattern,” “Aspect Sentence Pattern,” and “Selected pattern” based on the F-measure. The “Noun Phrase” is a commonly used part of speech such as a noun (NN), a proper noun (NNP) and a noun phrase and the most of aspects are noun-based phrases [48, 49]. The “Aspect Pattern” is patterns where aspect candidates are occurred. The “Aspect Sentence Pattern” is all extracted patterns through the morphological pattern recognizer. The “Selected Pattern” is some lengths patterns in Section 4.1. Table 24 shows the results of precisions, recalls and F-scores for each type of patterns. As shown in the table, the “Selected Patterns” with “LF” shows a higher accuracy (F-measure) for all three sources (80.86% - Movie Review, 63.07% - YouTube, 46.00% - Twitter) than the others. Thus, we decided to use the “Selected Pattern.”

To examine the performance of expression extraction, we compared all proposed methods, which are “Adj. Phrase,” “Expression Pattern,” “Expression Sentence Pattern,” and “Selected Pattern” based on the F-measure. The “Adj. Phrase” consists of adjective and verb phrases, because the most of expression are these types of phrases [48, 49]. The “Expression Pattern” is patterns where Expression candidates are occurred. The “Expression Sentence Pattern” is all extracted patterns through the morphological pattern recognizer. The “Selected Pattern” is some lengths patterns in Section 4.1. Table 25

shows the results of precisions, recalls and F-scores for each type of patterns. As shown in the table, the "Selected Pattern" shows a higher accuracy (F-score 58.01%) than the others. Thus, we decided to use the "Selected Pattern."

Table 24. The F-score of extracted aspect by the types of patterns

Data	Pattern	Extracted Aspect	Correct Aspect	Answer	Precision	Recall	F-Score
Movie Review	Noun Phrase	450	90	230	20.00%	39.13%	26.47%
	Aspect Pattern	4,999	230	230	4.60%	100.00%	8.80%
	Aspect Sentence Pattern	2,516	230	230	9.14%	100.00%	16.75%
	Selected Pattern	1,689	228	230	13.50%	99.13%	23.76%
	Aspect Sentence Pattern LF	1,068	227	230	21.25%	98.70%	34.98%
	Selected Pattern LF	329	226	230	68.69%	98.26%	80.86%
	Selected Pattern LF (frequency >= 1)	236	205	230	86.86%	89.13%	87.98%
	Co-occurrence	243	208	230	85.60%	90.43%	87.95%
YouTube	Noun Phrase	298	79	283	26.51%	27.92%	27.19%
	Aspect Pattern	3,354	188	283	5.61%	66.43%	10.34%
	Aspect Sentence Pattern	1,826	283	283	15.50%	100.00%	26.84%
	Selected Pattern	1,826	282	283	15.44%	99.65%	26.74%
	Aspect Sentence Pattern LF	400	195	283	48.75%	68.90%	57.10%
	Selected Pattern LF	589	275	283	46.69%	97.17%	63.07%
	Selected Pattern LF (frequency >= 1)	317	232	283	73.19%	81.98%	77.33%
	Co-occurrence	179	177	283	98.88%	62.54%	76.62%
Twitter	Noun Phrase	182	20	96	10.99%	20.83%	14.39%
	Aspect Pattern	3,965	70	96	1.77%	72.92%	3.45%
	Aspect Sentence Pattern	2,275	96	96	4.22%	100.00%	8.10%
	Selected Pattern	1,698	92	96	5.42%	95.83%	10.26%
	Aspect Sentence Pattern LF	262	73	96	27.86%	76.04%	40.78%
	Selected Pattern LF	304	92	96	30.26%	95.83%	46.00%
	Selected Pattern LF (frequency >= 1)	126	79	96	62.70%	82.29%	71.17%
	Co-occurrence	79	118	96	66.95%	82.29%	73.83%

Table 25. The F-score of extracted expressions by the types of patterns

Data	Pattern	Extracted Aspect	Correct Aspect	Answer	Precision	Recall	F-Score
Movie Review	Adj. Phrase	450	54	250	12.00%	21.60%	15.43%
	Expression Pattern	5,096	250	250	4.91%	100.00%	9.35%
	Expression Sentence Pattern	2,275	250	250	10.99%	100.00%	19.80%
	Selected Pattern	1,455	248	250	17.04%	99.20%	29.09%
	Expression Sentence Pattern LF	610	202	250	33.11%	80.80%	46.98%
	Selected Pattern LF	605	248	250	40.99%	99.20%	58.01%
	Selected Pattern LF (frequency >= 1)	329	220	250	66.87%	88.00%	75.99%
	Co-occurrence	283	206	250	72.79%	82.40%	77.30%
YouTube	Adj. Phrase	153	45	341	29.41%	13.20%	18.22%
	Expression Pattern	3,340	265	341	7.93%	77.71%	14.40%
	Expression Sentence Pattern	1,871	341	341	18.23%	100.00%	30.83%
	Selected Pattern	1,871	341	341	18.23%	100.00%	30.83%
	Expression Sentence Pattern LF	530	244	341	46.04%	71.55%	56.03%
	Selected Pattern LF	833	339	341	40.70%	99.41%	57.75%
	Selected Pattern LF (frequency >= 1)	449	278	341	61.92%	81.52%	70.38%
	Co-occurrence	424	290	341	68.40%	85.04%	75.82%
Twitter	Adj. Phrase	84	21	154	25.00%	13.64%	17.65%
	Expression Pattern	2,921	139	154	4.76%	90.26%	9.04%
	Expression Sentence Pattern	2,344	154	154	6.57%	100.00%	12.33%
	Selected Pattern	1,626	152	154	9.35%	98.70%	17.08%
	Expression Sentence Pattern LF	385	138	154	35.84%	89.61%	51.21%
	Selected Pattern LF	516	150	154	29.07%	97.40%	44.78%
	Selected Pattern LF (frequency >= 1)	258	110	154	42.64%	71.43%	53.40%
	Co-occurrence	283	206	250	72.79%	82.40%	77.30%

At this point, we examine how each type of pattern performs in terms of the processing time and the accuracy (F-measure). Figure 16 shows the results of extracting aspects and expressions from movie reviews. In terms of processing time, the selected pattern for extracting aspects (22,755 patterns used and 298 seconds spend) are about 5 times faster than all patterns used (59,111 patterns used and 1,491 seconds spend). Also, the selected patterns for extracting expressions (9,758 patterns used, 142 seconds spend)

are about 4 times faster than all patterns used (21,101 patterns used and 539 seconds spend). In addition, the selected patterns with the longest-first matching method ("Selected Pattern LF") showed the best performance on both the processing time and the accuracy (F-measure).

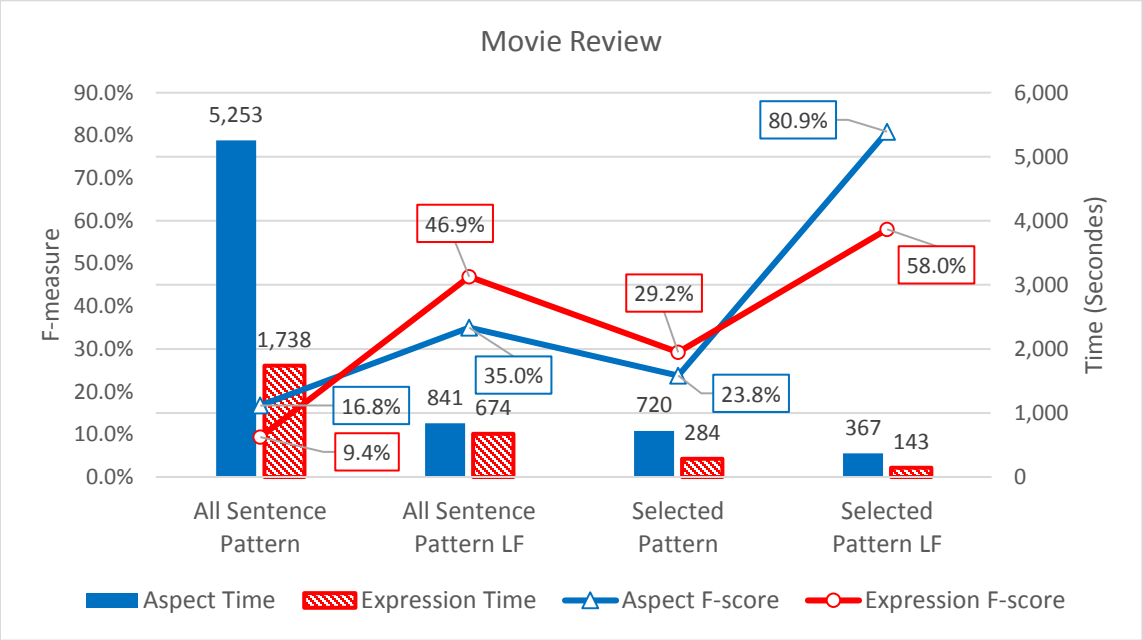

Figure 16. Results of extracting aspects and expressions by methods for movie reviews

Figure 17 shows the results of extracting aspects and expressions from YouTube comments. In terms of processing time, the selected pattern for extracting aspects (12,014 patterns used and 298 seconds spend) are about 6 times faster than all patterns used (71,178 patterns used and 1,801 seconds spend). Also, the selected patterns for extracting expressions (5,852 patterns used and 147 seconds spend) are about 9 times faster than all

patterns used (49,915 patterns used and 1,279 seconds spend). In addition, the selected patterns with the longest-first matching method ("Selected Pattern LF") showed the best performance on both the processing time and the accuracy (F-measure).

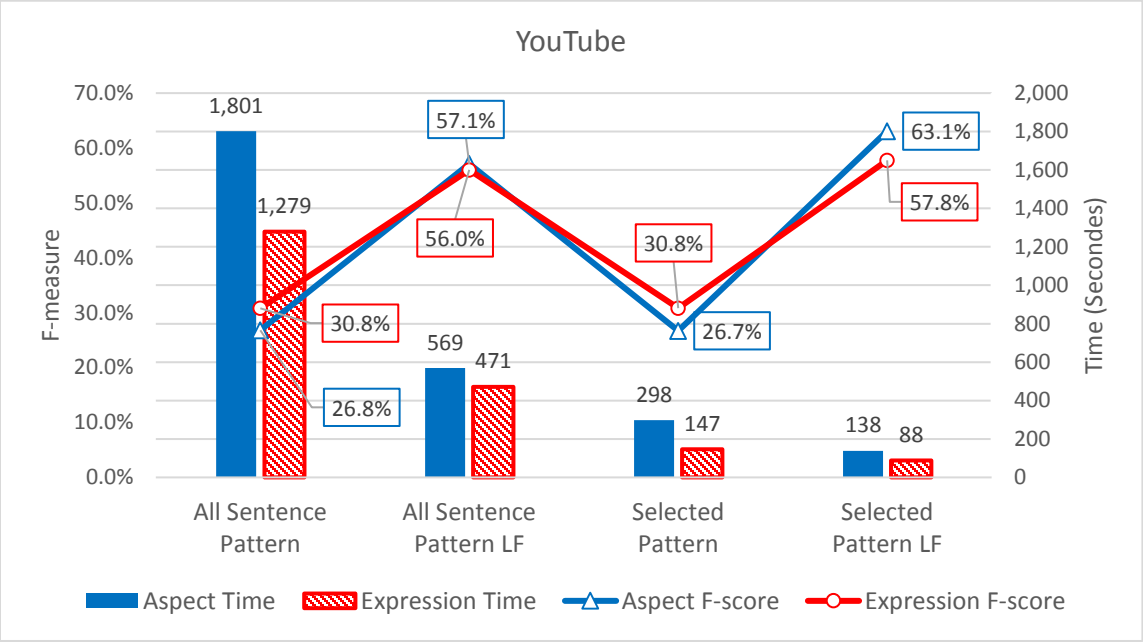

Figure 17. Results of extracting aspects and expressions by methods for YouTube

Figure 18 shows the results of extracting aspects and expressions from YouTube comments. In terms of processing time, the selected pattern for extracting aspects (14,893 patterns used, 378 seconds spend) are about 2 times faster than all patterns used (47,606 patterns used, 872 seconds spend). Also, the selected patterns for extracting expressions (8,627 patterns used, 221 seconds spend) are about 4 times faster than all patterns used (36,002 patterns used, 875 seconds spend). In addition, the selected patterns with the

longest-first matching method ("Selected Pattern LF") showed the best performance on both the processing time and the accuracy (F-measure).

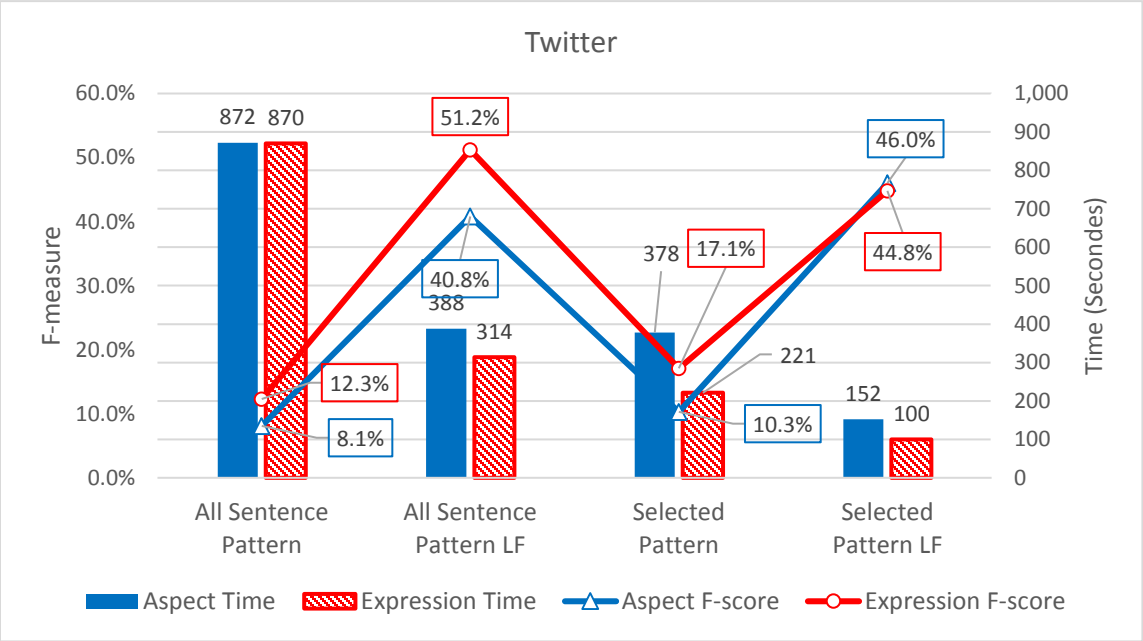

Figure 18. Results of extracting aspects and expressions by methods for Twitter

Through these results, we found that the 'Selected Pattern LF' showed the highest F-score (F-measure) and the shortest processing time for all data sources. It means that this method mostly affects both accuracy and processing time while "Selected Pattern" affects only the processing time. Therefore, we decided to use "Selected Pattern LF."

Even though, this method showed improved performances, the accuracies are still low. Therefore, we applied the frequency and co-occurrence of aspects and expressions as threshold for improving our method. Firstly, we used the frequency of aspects and

expressions as a threshold which is named as “frequency > 1.” After extracting aspects and expressions, the system filters out aspects or expressions when its frequency is one, because we assume that these aspects and expressions are meaningless. Secondly, the system retrieves pairs of aspects and expressions when these pairs are occurred in a sentence. Then, the system filters out aspects or expressions when its co-occurrence is one, because we assume that these are also meaningless.

As shown in the Figure 19, “Selected Pattern (frequency > 1)” method shows a higher F-score (88.0% for aspects and 88.0% for expressions) than “Selected Pattern LF” (80.9% for aspects and 58.0% for expressions) for Movie Review. “Co-occurrence” method shows a higher F-score (88.0% for aspect and 75.8% for expressions) than “Selected Pattern LF” method for Movie Review.

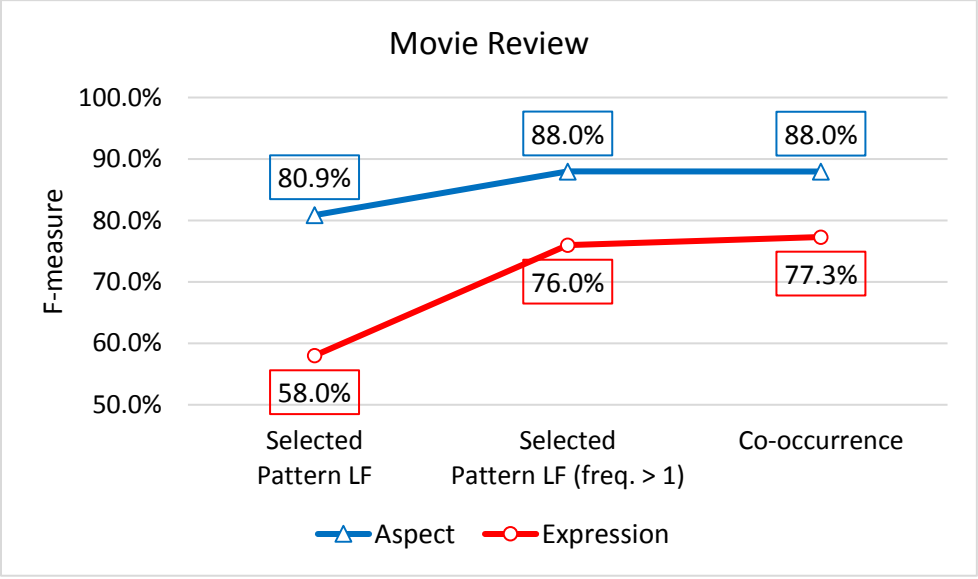

Figure 19. Results of extracting aspects and expressions with improving methods for movie review

As shown in the Figure 20, “Selected Pattern (frequency > 1)” method shows a higher F-score (77.3% for aspects and 76.6% for expressions) than “Selected Pattern LF” (63.1% for aspects and 57.8% for expressions) for YouTube. “Co-occurrence” method shows a higher F-score (76.6% for aspect and 75.8% for expressions) than “Selected Pattern LF” method for YouTube.

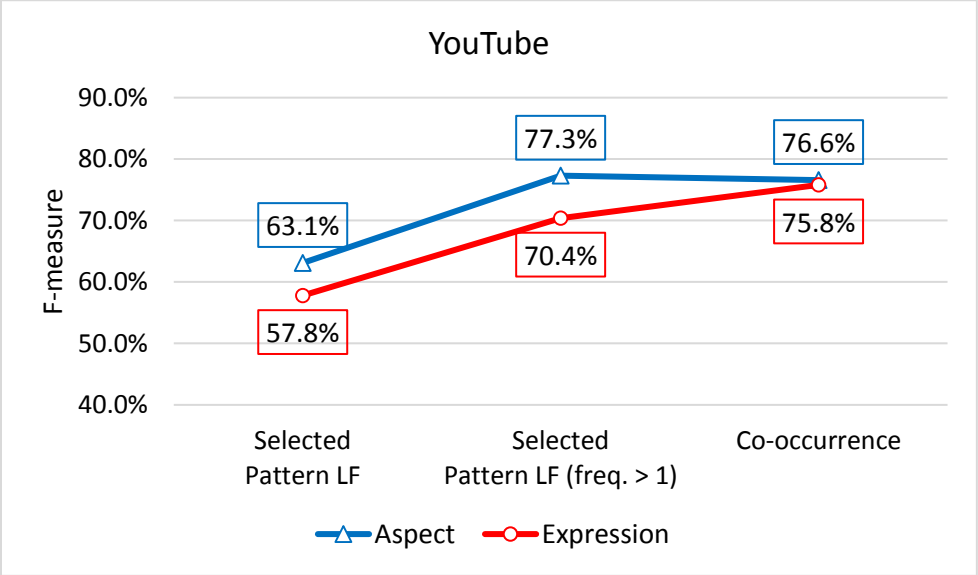

Figure 20. Results of extracting aspects and expressions with improving methods for YouTube

As shown in the Figure 21, “Selected Pattern (frequency > 1)” method shows a higher F-score (71.2% for aspects and 53.4% for expressions) than “Selected Pattern LF” (46.0% for aspects and 44.8% for expressions) for tweets. “Co-occurrence” method shows a higher F-score (73.8% for aspect and 77.3% for expressions) than “Selected

Pattern LF" method for tweets. This finding suggests that frequency and co-occurrence affects the F-score for extracting aspects and expressions.

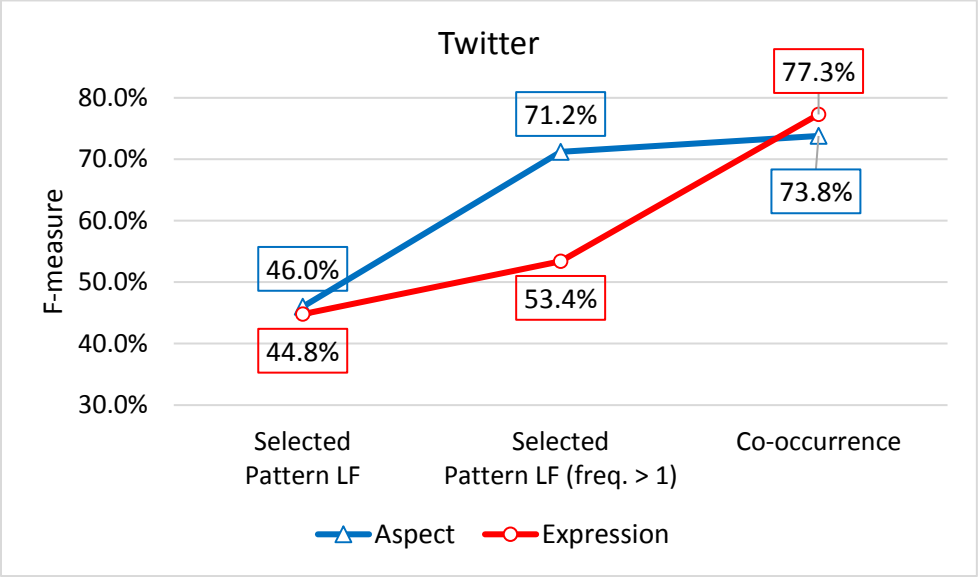

Figure 21. Results of extracting aspects and expressions with improving methods for Twitter

On the other hands, the Figure 22 and 23 show the results of precisions, recalls and F-scores depending on the numbers of co-occurrences. When the numbers of an aspect and an expression is greater than the average number of all co-occurrence which numbers are 227 for the aspects and 129 for the expressions, the precision of aspects is 100% (See Figure 22) and the precision of expression is 99.14% (See Figure 23). This finding suggests that user can adjust the threshold depending on the users' needs.

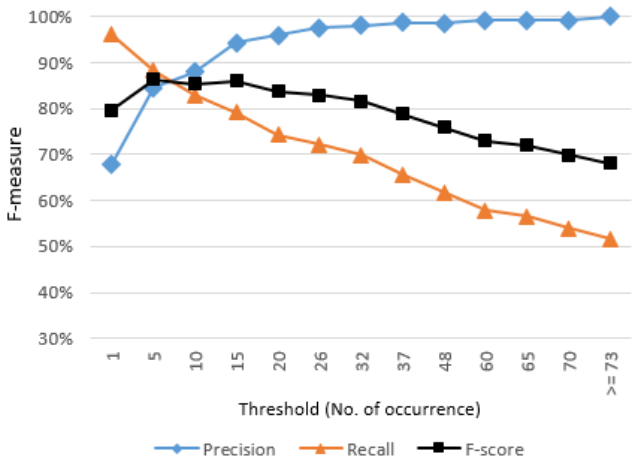

Figure 22. The F-score of extracted correct aspect by the length of patterns

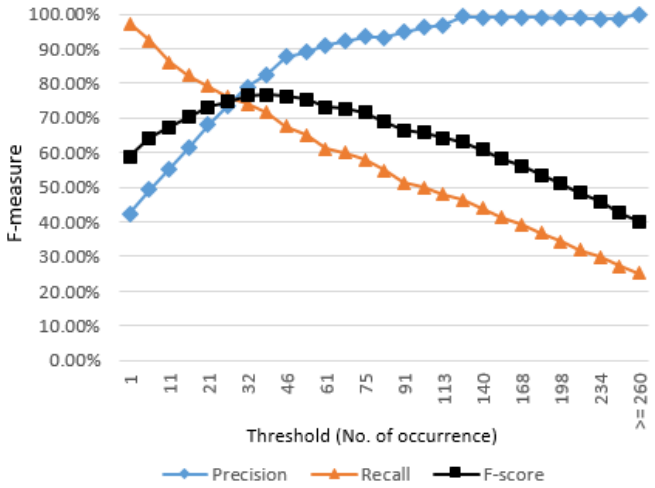

Figure 23. The F-score of extracted correct expression by the length of patterns

For a further possibility to use the patterns, we examined how the morphological sentence patterns work across domains, which are movie reviews, YouTube and Twitter. This cross-domain analysis is an experiment how the system extracts aspects and

expressions using patterns of each data source across other sources data. For example, the system extracts aspects and expressions from the YouTube comments and Twitter tweets using movie review's patterns as shown in Figure 24. Through this comparative cross-domain analysis, we discovered how the patterns are applicable for other source data.

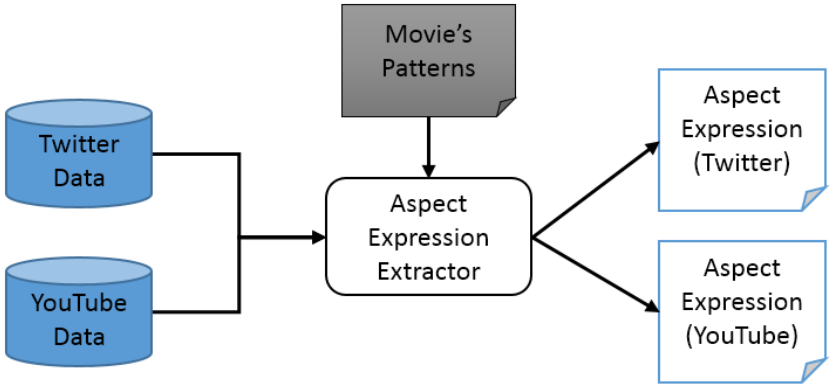

Figure 24. Cross-domain analysis

Table 26 and 27 show results of extracting aspects and expressions with cross-domain analysis. From all identified results, when we used same source of data and patterns, the F-score showed significantly higher score than the other patterns were used. We assumed that the result of YouTube and Twitter could give similar F-scores when we used the patterns of YouTube and Twitter are used each other because these are social media and having similar characteristics in terms of the length and number of POSs of a sentence. However, the results shown the similar F-score as other cross-domain analysis such as between patterns of movie review with YouTube comments (25.48%) and

patterns of YouTube comments with movie reviews (23.72%). This result implies that people express their opinion depending on each media.

Table 26. Results of cross-domain analysis for extracting aspects

Data	Pattern	Extracted Aspect	Correct Aspect	Answer	Precision	Recall	F-Score
Movie	Movie	**236**	**205**	**230**	**86.86%**	**89.13%**	**87.98%**
Movie	Twitter	238	70	230	29.41%	30.43%	29.91%
Movie	YouTube	296	67	230	22.64%	29.13%	25.48%
Twitter	Twitter	**126**	**79**	**96**	**62.70%**	**82.29%**	**71.17%**
Twitter	Movie	223	31	96	13.90%	32.29%	19.44%
Twitter	YouTube	255	33	96	12.94%	34.38%	18.80%
YouTube	YouTube	**309**	**231**	**283**	**74.76%**	**81.63%**	**78.04%**
YouTube	Movie	198	76	283	38.38%	26.86%	31.60%
YouTube	Twitter	168	66	283	39.29%	23.32%	29.27%

Table 27. Results of cross-domain analysis for extracting expressions

Data	Pattern	Extracted Expr	Correct Expr	Answer	Precision	Recall	F-Score
Movie	Movie	**283**	**206**	**250**	**72.79%**	**82.40%**	**77.30%**
Movie	Twitter	351	47	250	13.39%	18.80%	15.64%
Movie	YouTube	313	43	250	13.74%	17.20%	15.28%
Twitter	Twitter	**100**	**130**	**154**	**76.92%**	**64.94%**	**70.42%**
Twitter	YouTube	244	41	154	16.80%	26.62%	20.60%
Twitter	Movie	146	30	154	20.55%	19.48%	20.00%
YouTube	YouTube	**424**	**290**	**341**	**68.40%**	**85.04%**	**75.82%**
YouTube	Movie	207	65	341	31.40%	19.06%	23.72%
YouTube	Twitter	258	65	341	25.19%	19.06%	21.70%

Thus, we compared our method with existing methods, which are Hu and Liu [82], HashtagLex [83], Sentiment140Lex [83], and TS-Lex [84]. Hu and Liu is a traditional model with a relative small lexicon. HashtagLex, Sentiment140Lex and TS-Lex are sentiment lexicons for Twitter. As shown in the Table 28, our method yields a relatively higher F-score (78.13). Therefore, we suggest our model for building aspect-based lexicon for social media analysis.

Table 28. Comparison of F-score with related researches

Methods	F-Score	Multi Source	Social Media
HL[82]	60.49	No	No
HashtagLex [83]	65.30	No	Twitter
Sentiment140Lex [83]	72.51	No	Twitter
TS-Lex [84]	78.07	No	Twitter
Proposed Model	**78.13**	**Yes**	**Movie, Twitter, YouTube**

5.2. Sentiment Analyzer using MSP Model

In our previous research, we proposed a probability model [46]. It guarantees a relatively higher accuracy (89%) than other approaches and it can be used to analyze text-based social media data. However, there are some limitations on this model. The first one is that all words have polarity while the words are meaningless such as articles (“a” and “the”) and prepositions (“to” and “on”). This causes inaccurate results or over-analysis. Table 29 shows examples of the problem.

Table 29. Examples of limitation of probability model

Tweet	Positivity (%)
Freddie Gray Not Alone: 1997 Baltimore Police Case Raised Same Issue.#ColorOfHeart WakeUpAmerica	Positive (100)
RT @samswey: #FreddieGray isn't an isolated incident. Baltimore police dept killed 5 people last year. All were black. Every. Last. One.	Positive (87)

The second limitation is that this approach requires a human-coded process to build train-set to maintain the accuracy. This means that users must rebuild the train-set continuously because this method cannot analyze untrained-words such as new-words and it also requires a new train-set for a new topic [46]. Therefore, we propose an unsupervised and aspect-based sentiment analysis method using morphological sentence pattern (MSP) model to minimize the limitations [47, 48, 49]. We can expect a higher accuracy than our previous method without those limitations.

5.2.1. Method

The system consists of four main phases; collecting, extracting aspects and expressions, extracting sentiment patterns, and sentiment analysis. In the first phase, the collector crawls movie reviews, YouTube comments, and tweets. In the second phase, the aspect-expression extractor discovers aspects and expressions using the MSP model. In third phase, sentiment pattern extractor retrieve all candidate morphological sentence patterns for sentiment analysis. Finally, the sentiment analyzer matches the patterns with sentimental lexicon to the collected data. Figure 25 shows the system architecture and flow. In the rest of this section, we describe this method in detail.

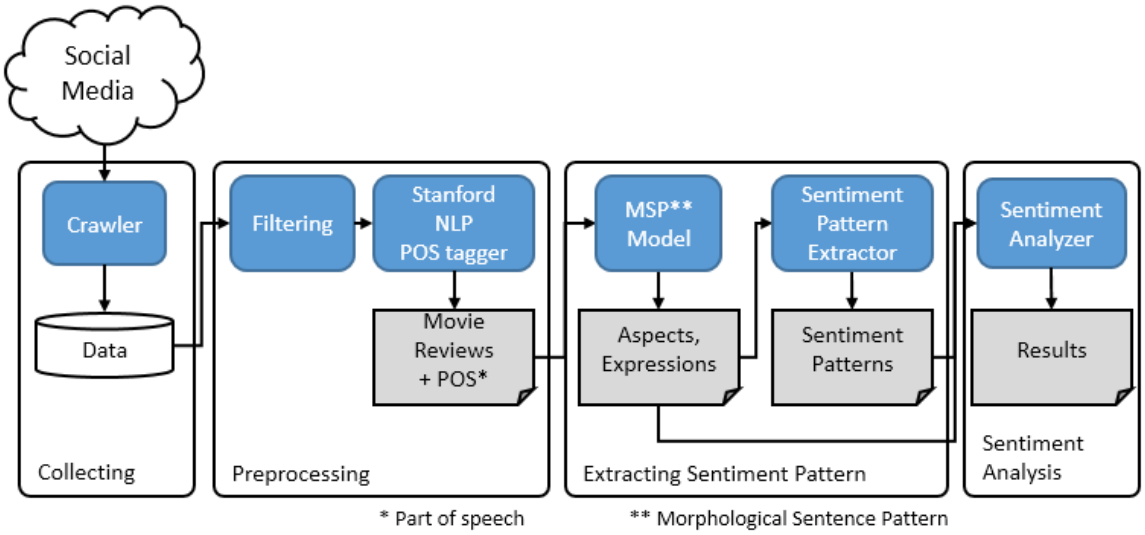

Figure 25. System architecture and flow

We developed a sentiment pattern extractor based on MSP model [10]. In our approach, we focused on which part of speech(s) exists between aspects and expressions. Table 30 shows a basic idea with examples. A pattern consists of three parts, which are

aspect, expression and Infix. The method matches aspects and expressions with sentimental lexicon when sentence contains the pattern. When matching the patterns, the system considers their sequence of POS(s) because it affects to their meaning [85].

Table 30. Examples of sentiment patterns

	Pattern			Text
	Aspect	*Infix*	*Expression*	
Words	character	-	best	the /DT/ best /JJS/ character /NN/
POS Pattern	NN	adjacent	JJS	
Words	Chris pratt	is	fine	chris /NNP/ pratt /NNP/ is /VBP/ fine /JJ/
POS Pattern	NNP+NNP	VBP	JJ	
Words	movie	is not	bad	this /DT/ movie /NN/ is /VBZ/ not /RB/ bad /JJ/
POS Pattern	NN	VBZ+RB	JJ	

To examine the patterns, we collected 200 documents for each source. Then, we randomly selected 100 matched results for each source and compared the results with a hand-coded answer-set labeled by two graduate students.

Table 31. Results of sentiment analysis on initial testing

	Movie Reviews	**YouTube Comments**	**Twitter Tweets**
Accuracy	87.5%	89%	83%
Partial Matching	8.5%	5%	10.5%
Mismatch	11%	1.5%	10%

At the initial testing, we just matched the patterns without any other processing. As shown in Table 31, the accuracy of the sentiment analysis is about 86.5 % on average. This number is reasonable accuracy compared with existing methods. However, we

found 2 main problems, which are partial matching problems and mismatching problems. The portions of the partial matching problems are about 8% and the mismatch problems are about 7.5% on average. Therefore, we decided to solve these problems for improving the accuracy.

Table 32. Examples of sentiment analysis results

Aspect : Expression	Sentiment	Correct	Text	Pattern
STORYLINE : GREAT	P	O	The action is intense, and the storyline is great.	(STORYLINE) /VBZ/ (GREAT)
ACTION : INTENSE	P	O	The action is intense, and the storyline is great.	(ACTION) /VBZ/ (INTENSE)
ACTING : GREAT	P	O	Movie also has great acting.	(GREAT) (ACTING)
MOVIE : GREAT	P	O	Movie also has great acting.	(MOVIE) /RB/ /VBZ/ (GREAT)
TREVORROW : GREAT	P	O	Colin trevorrow is a great director.	(TREVORROW) /VBZ/ /DT/ (GREAT)
SEQUEL : BANAL	N	O	Spielberg and company found a way to destroy its flavor with two banal sequels.	(BANAL) (SEQUEL)
MOVIE : SUCK	N	O	You'll most likely see title's following the words, "this movie sucked" or "i'm gonna go on a rant here".	(MOVIE) (SUCK)
ANYTHING : TERRIBLY	N	X	To be honest there wasn't anything terribly new in this film,	(ANYTHING) (TERRIBLY)
EFFECT : SPECIAL	P	X	This movie has great special effects.	(SPECIAL) (EFFECT)
SPECIAL EFFECT : GREAT	P	O	This movie has great special effects.	(GREAT) (SPECIAL EFFECT)
CHARACTER : LIKABLE	P	O	Sure chris pratt's character is the only likable person in the film but the dinosaurs are epic.	(CHARACTER) /VBZ/ /DT/ /JJ/ (LIKABLE)
ACTING : BELIEVABLE	P	O	Aside from a few overdramatic scenes by claire, the acting is really believable.	(ACTING) /VBZ/ /RB/ (BELIEVABLE)

Table 32 shows examples of sentiment analysis results of the initial testing. In this table, the ninth and tenth rows indicate the partial matching problem. In this example, the system extracted both ‘EFFECT’ and ‘SPECIAL EFFECT’ as aspects from a same sentence because the system matched all possible patterns even though the ‘EFFECT’ is a part of ‘SPECIAL EFFECT’. The portion of the problem is about 8% on average. To solve this problem, we applied a priority to the longest-matched word(s) because “SPECIAL EFFECT” seems more meaningful as an aspect in the example sentence, “THIS MOVIE HAS GREAT SPECIAL EFFECTS.”

In addition, the third and fourth rows indicate the mismatching problem. In this example, the analyzer extracted an expression ‘GREAT’ with ‘ACTING’ and ‘MOVIE’ in a same sentence because the system also matched all possible patterns. The portion of the problem is about 7.5%. To solve this problem, we applied a priority to the nearest pair because the expression, “GREAT” targets “ACTING” in the example sentence, “MOVIE ALSO HAS GREAT ACTING.”

5.2.2. Experiment Result

To experiment our method, we collected 1,000 documents from movie review sites (IMDb, Rotten Tomatoes, and Metacritic), Twitter and YouTube related to a movie, 'Jurassic world'. Then, the pattern extractor generated the sentiment patterns using the aspects and expressions, which are extracted from the MSP model. From all extracted patterns, we selected 165 patterns for movie reviews, 113 patterns for tweets, and 102 patterns for YouTube comments when their frequency was 2 or greater.

102 patterns from YouTube

YouTube	count	%
None(adjacent)	121	36.1%
DT	45	13.4%
VBD	26	7.8%
IN	11	3.3%
IN DT	8	2.4%
VBZ	8	2.4%
VBD RB	6	1.8%
-LRB-	4	1.2%
VBZ RB	4	1.2%
"	3	0.9%
:	3	0.9%
JJ	2	0.6%
MD VB DT	2	0.6%
PRP VBZ RB	2	0.6%
RB	2	0.6%

165 patterns from movie

Movie	count	%
None(adjacent)	165	35.6%
DT	21	4.5%
VBZ	19	4.1%
IN	16	3.5%
VBD	11	2.4%
VBZ RB	10	2.2%
IN DT	9	1.9%
RB	9	1.9%
VBZ DT	8	1.7%
VBP	6	1.3%
" VBZ DT	4	0.9%
NN	4	0.9%
JJ	3	0.7%
MD VB	3	0.7%
POS	3	0.7%

113 patterns from tweets

YouTube	count	%
None(adjacent)	66	26.6%
NNP	37	14.9%
JJ	12	4.8%
RB	10	4.0%
DT	8	3.2%
VBZ	6	2.4%
IN	4	1.6%
NNP NNP	4	1.6%
IN DT	3	1.2%
TO VB	3	1.2%
VBG	3	1.2%
CD	2	0.8%
DT JJ	2	0.8%
DT NN	2	0.8%
IN NNP NN VBZ	2	0.8%

Figure 26. Example of extracted sentiment patterns

Figure 26 shows examples of extracted sentence patterns in order of their frequency. In this result, about 30% of aspects and expressions are adjacent (distance is 0) and most others have one or two POS(s) (distance is 1 or 2) between aspects and expressions. This means most of the expressions affect to near aspects on all three

sources. On the other hands, we found a characteristic from extracted sentence patterns. We calculated a distance of aspects and expressions (word proximities) from all extracted patterns as shown in Figure 27. The average distance is 1.17 (movie reviews: 1.81, YouTube: 1.41, Twitter: 1.91). This means that most of the extracted patterns have one or two part of speech(s) between aspects and expressions on average.

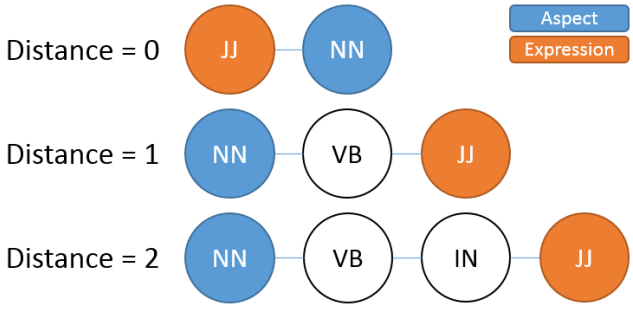

Figure 27. Calculating word proximity between aspect and expression

To examine the sentiment analyzer, we used a sentiment lexicon generated by Bing Liu [117], which contains 2,003 positive words and 4,782 negative words. In addition, we added 102 positive words and 98 negative words extracted from the MSP model. Also, we used 102 stop-words including reserved words by service providers such as “HTTP,” “RT,” and “@.” Then, we analyzed 1,000 document collected from each data source. Then, we also randomly selected 200 results for each data source and compared the results with hand-coded answer-set labeled by two graduate students same as in the initial testing.

Table 33. Results of sentiment analysis

	Accuracy
Movie Reviews	92.5%
YouTube Comments	93%
Twitter Tweets	88%
Average	91.2%

As shown in Table 33, the accuracy of the sentiment analysis was at the level of about 91.2% with the proposed MSP model, which improved the partial matching problem and the mismatching problem. The number is improved by 4.7% from the results of the initial testing from 86.5 %. Thus, this model improved the accuracy (91.2%) than existing approaches as shown in Table 34.

Table 34. Comparison with existing approaches

Method	Accuracy
PANAS-t [86]	0.677
Emoticons [87]	0.817
SASA [88]	0.649
SenticNet [89]	0.590
SentiWordNet [90]	0.643
SentiStrength [91]	0.639
Happiness Index [92]	0.639
LIWC [93]	0.815
Probability Model [46]	0.890
Our approach	**0.912**

6. AN EXTRACTING METHOD OF MOVIE GENRE SIMILARITY USING ASPECT-BASED APPROACH

The movie recommendation is a manner of advertisement or promotion that targets customers in the movie industry. In this section, we propose an extracting method of movie genre similarity for movie recommendation using the MSP model and machine learning algorithms. Our method consists of two main methods, which are "TDF-IDF" and "Genre Score." The "TDF-IDF" is designed to find genre-representative keywords. The "Genre Score" is designed to discover a similarity of movies with consideration for genres using the keywords. To find similar movies, we used K-means and K-NN algorithms based on the similarity.

6.1. Method

Figure 28 shows the system architecture and flow. This system consists of four main phases; collecting data, extracting aspects and expressions, calculating genre scores, and recommendation. At the first phase, the crawler collects movie reviews and YouTube comments using movie names as seed keywords such as "Avengers: Age of Ultron," "Deadpool," "Jurassic World," and "Star Wars: The Force Awakens." Then, the system extracts aspects and expressions from the collected data using the MSP model [10]. Using "TDF-IDF" method, the system calculates an importance of aspects and expressions by

each genre. These aspects and expressions are used as feature keywords to calculate a genre similarity. The system then calculates the "Genre Score," which is developed to find correlations between genres and movies. The score indicates a degree of relations between four representative genres, which are "action," "animation," "comedy," and "horror." Then, this system recommends movies based on the results of K-means and KNN algorithms using R Studio [100, 101, 102]. In the rest of this section, we describe this method in-detail.

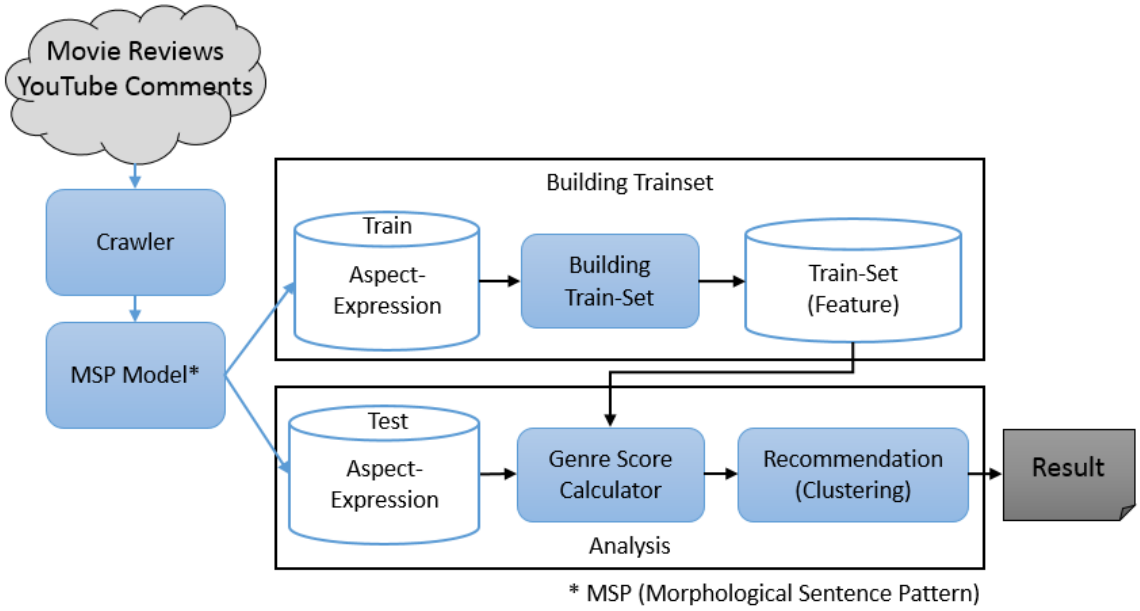

Figure 28. System architecture and flow

First of all, the system extracts aspects and expressions using the MSP model [10, 11]. Then, the system retrieves genre representative aspects and expressions based on the TDF-IDF. This method helps to filter out commonly used aspects and expressions, which

occur frequently in every genres. To calculate genre similarity, the system selects the top 100 aspects and expressions as features.

At the beginning of this research, we tried to use the TF-IDF to discover genre representative keywords because TF-IDF is broadly used to calculate an importance of a keyword [55, 56]. However, this method was not sufficient to use for this research. For example, when a keyword has a higher TF-IDF score, the keyword is might be a movie-specific word such as a character name, actor or location, rather than a genre-representative keyword.

$$tdf_{t,d} = (1 + \log(tf_{t,d})) \times \boldsymbol{df_{t \in dt}} \qquad (5)$$

$$\boldsymbol{tdfidf_{t,d}} = tdf_{t,d} \times \log\left(\frac{N}{df_t}\right) \qquad (6)$$

$$\boldsymbol{Feature} = Top_{100}(tdfidf_{t,d}) \qquad (7)$$

Therefore, we proposed a method named "TDF-IDF" based on the TF-IDF. In this method, the TDF is the weighted TF with a document frequency of all relevant movies (5). An aspect can be a genre or a group of movies' representative aspect chosen by the TDF. Then, the TDF is multiplied by IDF as shown in the equation (6). After calculating TDF-IDF, we selects the Top 100 aspects and expressions as features to calculate the genre score as shown in the equation (7).

$$G_p = \sum_{t \in Tp} tfidf(t) \quad (8)$$

$$G_q = \sum_{t \in Tq} tfidf(t) \quad (9)$$

$$\boldsymbol{Genre\ Score} = (G_p - G_q) \div (G_p + G_q) \quad (10)$$

$$\boldsymbol{Genre\ Score_{all}} = Genre\ Score_{aspect} + Genre\ Score_{expression} \quad (11)$$

The genre score is designed to discover a correlation between a target movie and movie genres using all extracted features based on the TDF-IDF scores. As shown in the equation (8), G_p is all the accumulated TDF-IDF scores when a set of target movie's aspects t appears in T_p, which is a set of Top 100 aspects are extracted from a relevant genre's reviews. As shown in the equation (9), G_q is all of the accumulated TDF-IDF scores when a set of target movie's aspects t appears in T_q. This is the top 100 aspects extracted from other genre's reviews. Therefore, G_p is how often a target movie's aspect occurs in relevant movie reviews and the G_q is how often a target movie's aspect occurs in irrelevant movie reviews. Finally, the system calculates a gap of $G_p$and G_q as a degree of correlation between the relevant movie and the irrelevant movies (10) as the genre score. The analyzer calculates the genre score by aspects and expressions separately since some movies have more than one genre and could not be categorized into either one of them. In section 6.2, we examined why both aspects and expressions should be used in

this research based on an experiment. Finally, the $\boldsymbol{Genre\ Score_{all}}$ is calculated as shown in equation (11).

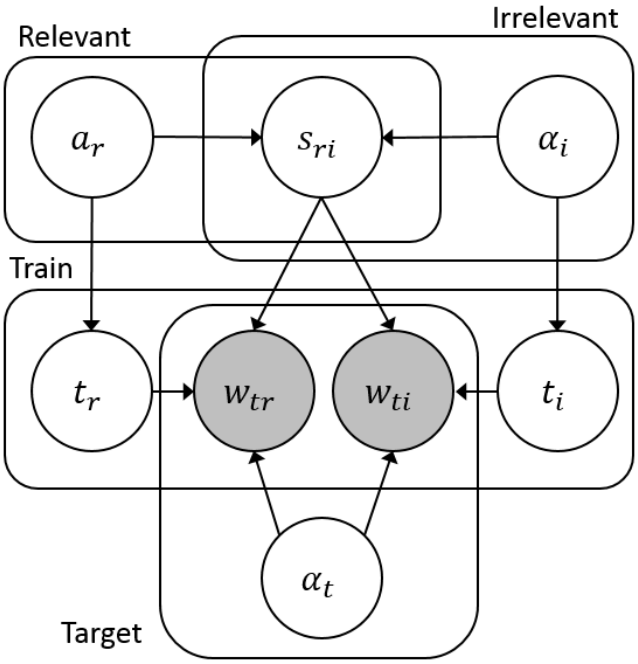

Figure 29. Factorized method by data-sets

Figure 29 shows our factorized method in terms of data-sets and flows. A given a is extracted aspects and expressions from relevant movie reviews r, irrelevant movie reviews i and target movie reviews t. Therefore, a_r is a set of aspects from relevant movies, a_i is a set of aspects from irrelevant movies, and a_t is a set of aspects from target movies. $s_{ri}$is a set of intersection between a_r and a_i. We used this set as stopwords to filter out commonly used aspects. t_r and t_i are a set of $a_r - a_i$ and $a_i - a_r$. These two sets are representatives of relevant movies and

irrelevant movies. Finally, the system extracts W_{tr}, which is a set of intersection between t_r and a_t , and W_{ti}, which is a set of intersection between t_i and a_t. W_{tr} is a set of representative aspects from relevant movies and W_{ti} is a set of representative aspects from irrelevant movies. The numbers of these two sets would be used to calculate the genre score.

6.2. Experiment Result

To verify our method, we selected the Top 100 movies listed on the box office mojo (http://www.boxofficemojo.com) released from January 2015 to May 2016. Then, we divided the 100 movies into 3 groups; 16 movies for train-set, 4 movies for test-set, and the remaining 80 movies for recommendation.

Table 35. Selected movies for extracting genre specific features

Group	Movies	Labeled Genres	Rank
Action	Star Wars: The Force Awakens	Action, Adventure, Fantasy	1st
	Jurassic World	Action, Adventure, Sci-Fi	2nd
	Avengers: Age of Ultron	Action, Adventure, Sci-Fi	3rd
	Deadpool	Action, Adventure, Comedy	4th
Animation	Inside Out	Animation, Adventure, Comedy	5th
	Minions	Animation, Comedy, Family	8th
	Zootopia	Animation, Action, Adventure	10th
	Home	Animation, Adventure, Comedy	20th
Comedy	Daddy's Home	Comedy	27th
	Trainwreck	Comedy, Romance	34th
	Get Hard	Comedy, Crime	38th
	Sisters	Comedy	41st
Horror	10 Cloverfield Lane	Drama, Horror, Mystery	50th
	Insidious:Chapter 3	Fantasy, Horror, Thriller	71st
	Poltergeist (2015)	Horror, Thriller	72nd
	The Boy	Horror, Mystery, Thriller	89th

As shown in the Table 35, we further categorize 16 movies into 4 movie genres, which were “action,” “animation,” “comedy,” and “horror” for the train-set. These movies are generally co-labeled with other genres. For example, action movies are

labeled with the “adventure,” “fantasy,” or “Sci-fi,” animation movies are labeled with “family,” comedy movies are labeled with “drama,” and horror movies are labeled with “thriller,” or “mystery.” We assumed that these movies were representative movies for each selected genre. Our system calculated the genre scores based on these groups.

Action		Animation		Comedy		Horror		
Keyword	Score	Keyword	Score	Keyword	Score	Keyword	Score	
HYPE	18.86	CUTE	28.72	BOX	26.32	MONSTER	33.25	Aspect
CIVIL	17.99	CARTOON	27.98	CENA	26.21	SCARE	31.21	
WAR	17.33	MINION	22.03	DAD	25.53	JUMP	26.48	
SIDE	16.99	TOY	21.66	STEP	25.53	TBH	26.32	
DARK	14.94	ANIMATION	17.12	COMEDIAN	22.16	GHOST	26.18	
WARS	14.77	PIXAR	15.94	BALL	22.16	DOLL	25.71	
SUPPOSE	14.48	LANGUAGE	15.15	AMY	21.06	JUMPSCARE	24.03	
COMPARE	14.17	MESSAGE	14.68	WAHLBERG	20.43	THRILLER	23.61	
CREATE	13.60	HERO	14.68	PARTY	20.36	PRODUCER	23.15	
KYLO	13.40	BLUE	14.48	STUPID	20.29	ZOMBIE	21.66	
CENTURY	27.59	ANIMATED	19.09	COMEDY	33.21	HOUSE	32.75	Expression
CHILL	26.48	MESSAGE	16.99	JOHN	24.94	HORROR	31.21	
BATTLE	25.37	EMOTIONAL	15.84	FUNNIEST	23.95	SCARIEST	30.04	
ORDER	22.16	REFERENCE	14.48	UNFUNNY	21.01	SCARY	21.99	
DEADPOOL	21.40	SHORT	13.41	WAHLBERG	20.36	TRAP	21.83	
FIGHT	20.71	BUSY	13.40	WILL	19.65	SCARE	20.06	
FORCE	20.07	THEORY	12.16	UGLY	18.16	SERIES	19.65	
FRANCHISE	19.65	CAT	12.16	TOP	17.99	PSYCHO	18.86	
BADASS	17.99	FROZEN	11.39	MARK	17.12	KILLER	18.86	
AWAKENS	17.37	MIXED	11.39	HAHAHA	16.99	LOUD	18.86	

Figure 30. Examples of feature words by genres from movie reviews

To build train-sets, we collected 100 movie reviews for each movie from IMDb, Rotten Tomatoes, and Metacritic. Then, the system extracted aspects and expressions using the MSP model. Figure 30 shows examples of extracted aspects and expressions with the TDF-IDF scores by the genres from movie reviews. Then, the analyzer

calculated the genre score using the top 100 aspects and expressions as features for each movie.

Figure 31 shows the results of calculated genre scores to see how the results are distinct between own genre and other genres based on the distances of genre scores. In this figure, the gray line shows the distance between the results of own genres and other genres. In these results, horror is the most distinguishable unlike comedy. This means that the features of comedy are a more commonly used in movie reviews.

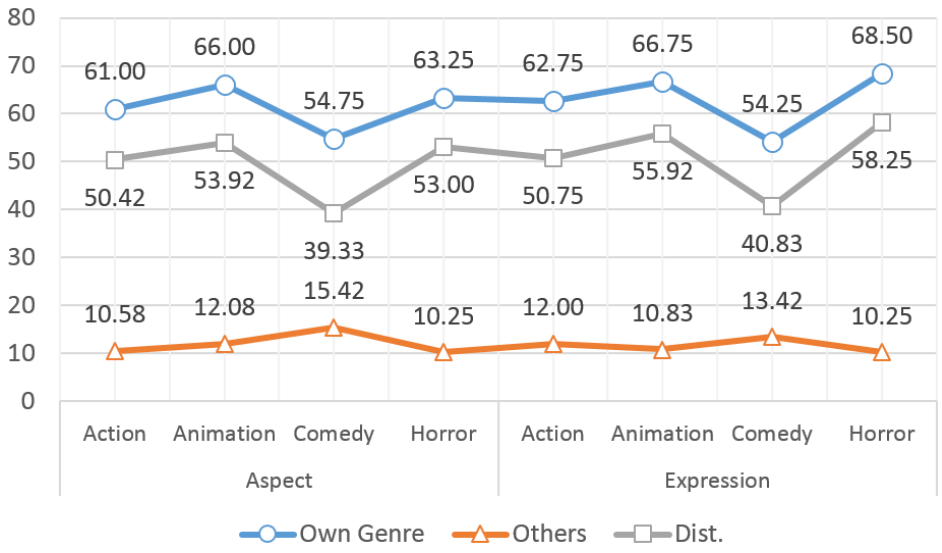

Figure 31. Results of the genre scores between own genres and other genres from movie reviews

We also collected 2,000 comments from YouTube on official movie trailers for each movie to build train-sets. Then, the system extracted aspects and expressions using the MSP model. Figure 32 shows examples of extracted aspects and expressions from YouTube comments with the TDF-IDF score by the genres. The TDF-IDF score indicates

a degree of importance of a genre. Based on the score, we selected the top 100 aspects and expressions as feature keywords. Then, the system calculated the genre score for each movie using the keywords.

Action		Animation		Comedy		Horror		
Keyword	**Score**	**Keyword**	**Score**	**Keyword**	**Score**	**Keyword**	**Score**	
SOLO	64.61	PIXAR	42.69	FERRELL	39.00	GHOST	43.51	Aspect
ULTRON	63.12	ANIMATION	36.74	SNL	29.40	POLTERGEIST	24.39	
GALAXY	62.43	ANIMATED	31.93	AMY	25.22	PARANORMAL	23.15	
SAGA	62.33	MINION	30.61	FEY	24.20	SCARE	21.37	
MARVEL	58.28	COP	24.90	POEHLER	21.62	JUMP	20.72	
HULK	56.58	DISNEY	23.23	CENA	21.40	JUMP-SCARE	20.36	
AWAKENS	54.07	RABBIT	23.23	BELLY	18.98	CREEPY	18.98	
AVENGER	51.74	DINOSAUR	22.67	TRAINWRECK	18.68	DEMON	18.68	
SOLDIER	50.71	SMITH	20.01	MARTIN	17.99	FRIGHT	18.29	
REBEL	46.53	TAKEAWAY	19.65	BASKETBALL	17.85	SCARY	17.71	
TRILOGY	70.50	POLICE	34.05	RAUNCHY	31.34	SPIRITUAL	26.18	Expression
SUPERHERO	60.65	DISTINCTIVE	31.75	VULGAR	24.82	HORROR	25.77	
BLOCKBUSTER	55.82	IMAGINATIVE	28.05	PAIRING	17.99	CREEP	25.37	
CORPORATE	50.24	ANIMATION	24.49	COMEDIAN	17.62	SPOOKY	24.82	
MILITARY	48.06	DELIGHTFUL	22.62	GAG	17.41	SUPERNATURAL	23.62	
THREAT	47.21	COLOURFUL	21.75	STAND-UP	16.26	INSIDIOUS	23.46	
ANTICIPATED	47.21	SLY	20.69	GAY	15.94	SCARIEST	22.67	
DESTRUCTION	46.87	FOX	19.55	AFFECTION	15.84	DOLL	21.83	
BLOODY	45.27	CROWN	19.39	NUDITY	15.84	PARANORMAL	21.62	
AWE	44.71	YELLOW	18.27	SEX	15.39	EERIE	20.06	

Figure 32. Examples of feature words from YouTube comments

Also, Figure 33 shows the results of calculated genre scores between its own genre and other genres as the result of movie reviews. In this case, comedy and horror are most clearly distinguishable unlike animation, meaning that the features of animation are more commonly used in YouTube comments.

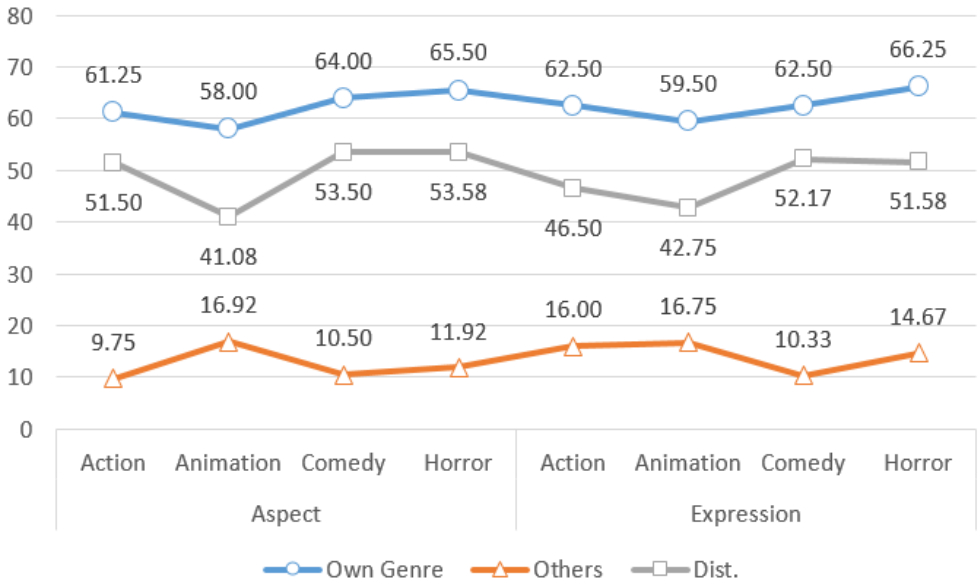

Figure 33. Results of the genre scores between own genres and other genres from YouTube comments

To build test-sets, we selected a movie for each genre, which is "Batman v Superman: Dawn of Justice" for action, "Ted 2" for comedy, "The Peanuts Movie" for animation, and "The Visit" for horror. Table 36 and Table 37 show results of genre scores extracted from movie reviews and YouTube comments based on the four genres by the aspect and expression. In these results, a movie showed the highest genre score in own genre. For example, an action movie had the highest genre score in the action genre. However, in the case of an animation movie ("The Peanuts Movie"), both the result of aspects and expressions are not distinct from other genres. This means that people use more common aspects and expressions about animation. Figure 34 shows the results of the normalized genre score based on percentages (%). In these results, all movies are well categorized into their genres.

Table 36. Results of genre score from movie reviews

Movie	Genre	Aspect	Expression	Total
Batman v Superman (Action)	**Action**	**35**	**32**	**67**
	Animation	8	-35	-27
	Comedy	-45	-6	-51
	Horror	-2	-8	-10
The Peanuts Movie (Animation)	Action	17	4	21
	Animation	**51**	**50**	**101**
	Comedy	-8	-2	-10
	Horror	-16	-25	-41
Ted 2 (Comedy)	Action	-11	10	-1
	Animation	58	-6	52
	Comedy	**15**	**68**	**83**
	Horror	-30	1	-29
The Visit (Horror)	Action	-36	19	-17
	Animation	25	-65	-40
	Comedy	-6	-6	-12
	Horror	**61**	**90**	**151**

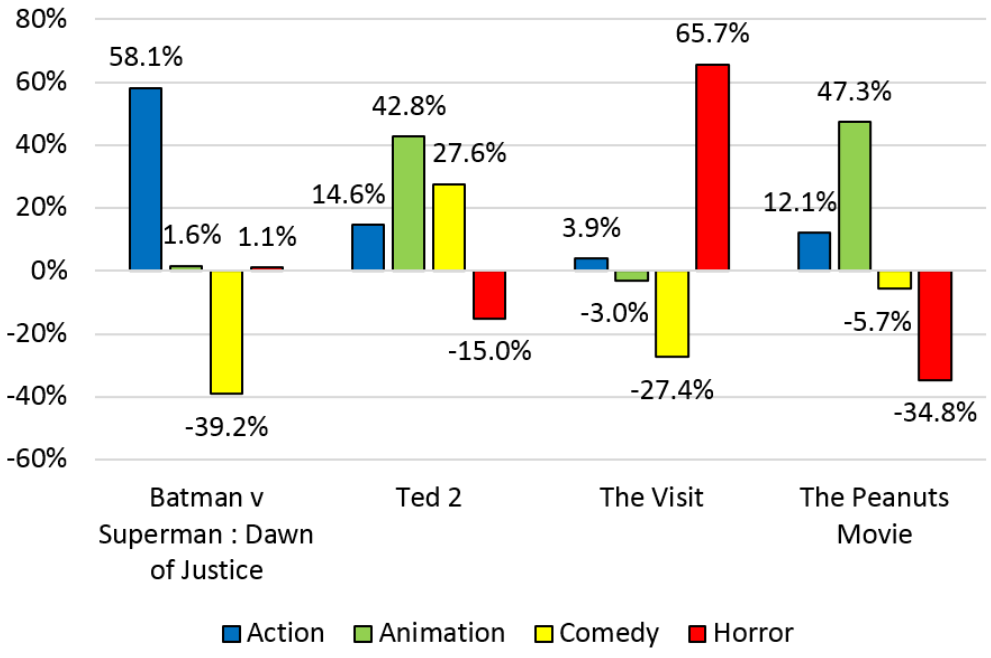

Figure 34. Results of genre score (%) for test movies using movie reviews

On the other hand, the genre score of “comedy” in YouTube comments was a greater than the result of movie reviews and the genre score of “animation” in movie reviews is greater than the result of YouTube comments. This difference came from how the people express their opinions depending on the media. Thus, we verified our method with both movie reviews and YouTube comments (See Table 36 and Table 37).

Table 37. Results of genre score from YouTube comments

Movie	Genre	Aspect	Expression	Total
Batman v Superman (Action)	**Action**	**54**	**66**	**120**
	Animation	-15	-9	-24
	Comedy	-2	-18	-20
	Horror	18	-3	15
The Peanuts Movie (Animation)	Action	6	9	15
	Animation	**15**	**11**	**26**
	Comedy	-33	25	-8
	Horror	-13	-61	-74
Ted 2 (Comedy)	Action	-5	31	26
	Animation	4	-27	-23
	Comedy	**37**	**42**	**79**
	Horror	9	-25	-16
The Visit (Horror)	Action	-19	-7	-26
	Animation	16	-15	1
	Comedy	-28	0	-28
	Horror	**56**	**40**	**96**

Figure 35 shows the results of the normalized genre score based on percentages (%). In these results, all movies are well categorized into their genres like the results of movie reviews. The system recommends relevant movies based on these results.

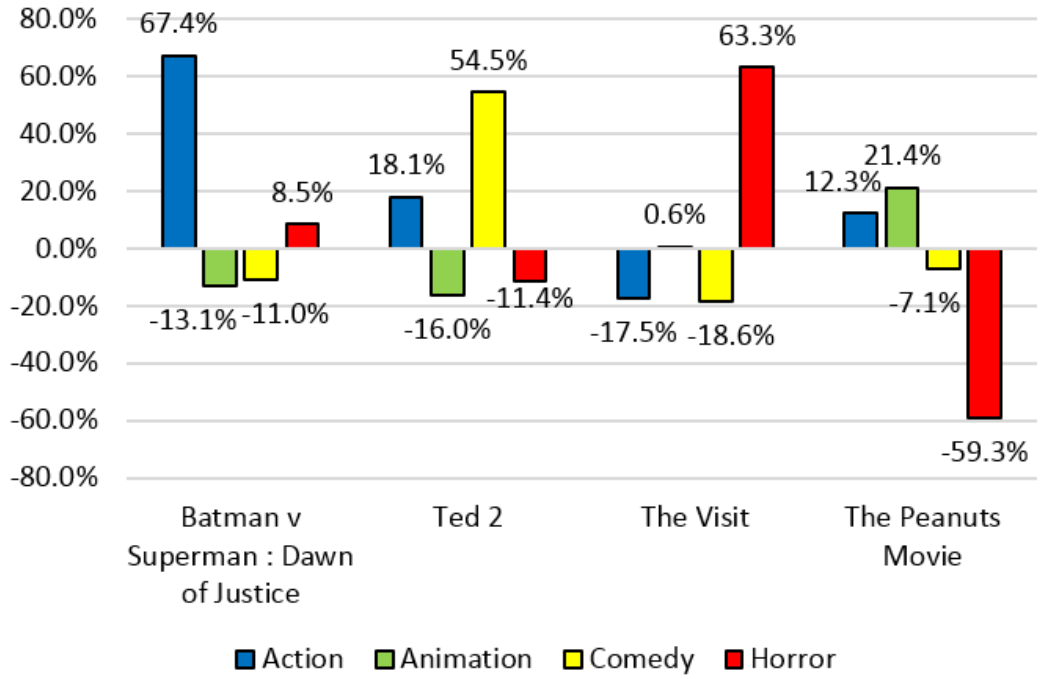

Figure 35. Results of genre score (%) for test movies using YouTube comments

As described in Section 2, we used both the K-Means and the K-NN algorithms to find similar movies to recommend. Firstly, we use K-Means clustering algorithm to group 80 to find similar movies using the distance of genre scores. The system recommends movies based on the results. We selected two k numbers, which were $k = 9$ (69.7% of variance) and $k = 11$ (82.1% of variance) based on the elbow method [62]. The system firstly finds relevant movies based on $k = 11$ because it seemed a stronger relation between movies in a group than the results of $k = 9$. Using R-Studio, we calculated and visualized the clustering results as shown in Figure 36 and 37. In these figures, we labeled the target movies (test-sets) and their groups. Table 38 and 39 show relevant movies by each target movie based on the K-Means clustering results. Figure 36

and Table 38 show the results of movie reviews, and Figure 37 and Table 39 show the results of YouTube comments.

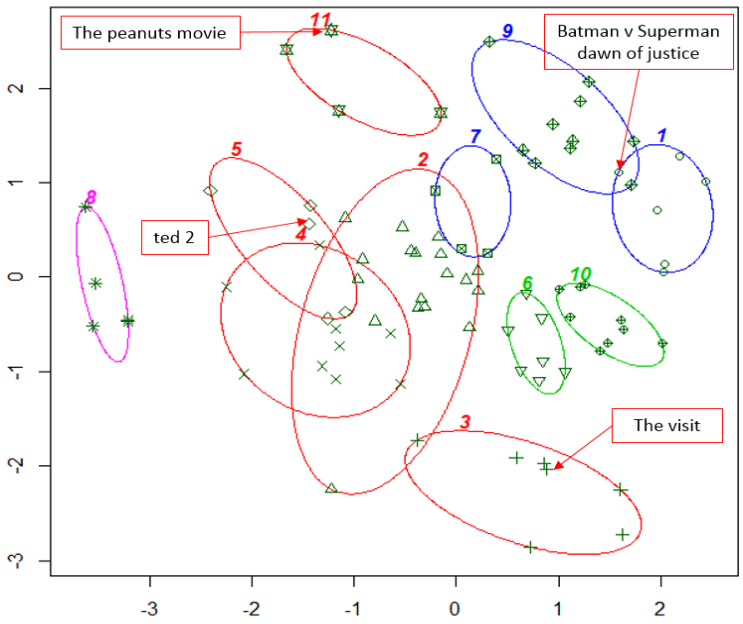

Figure 36. Results of K-Means clustering using movie reviews (K=11)

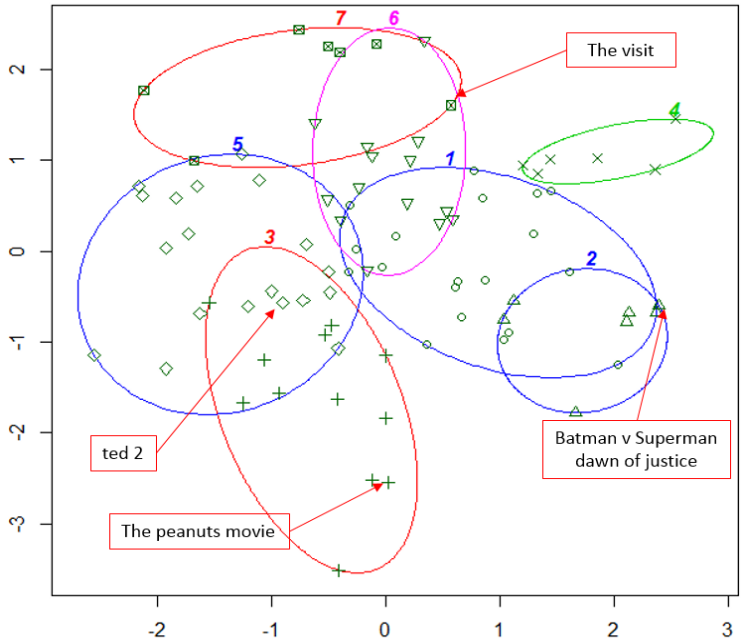

Figure 37. Results of K-Means clustering using YouTube comments (K=11)

Table 38. Movie recommendations by K-means result using movie reviews

Source Movie	1st Group, K=11 (Genre)	2nd Group, K=7 (Genre)	
Batman v Superman: Dawn of Justice (AC, AD, SF, F)	• 13 Hours: The Secret Soldiers of Benghazi (AC, D, T) • American Sniper (AC, AD, D) • **Fantastic Four (AC, AD, SF)** • London Has Fallen (AC, AD, CR, D) • **The Divergent Series: Allegiant (AC, AD, SF, R, M)**	• Ant-Man (AC, AD, C, SF, T) • Furious 7 (AC, AD, CR, T) • Jupiter Ascending (AC, AD, SF, F) • **Pan (AC, AD, FM, F)** • **Terminator Genisys (AC, AD, SF)** • The Huntsman: Winter's War (AC, AD, D, F) • The Jungle Book (AC, AD, D, FM) • The Martian (AC, AD, SF, D) • Tomorrowland (AC, AD, SF, FM)	
The Peanuts Movie (AN, AD, C, FM)	• Alvin and the Chipmunks: The Road Chip (AN, AD, C, FM) • Paddington (AN, C, FM) • The SpongeBob Movie: Sponge Out of Water (AN, AD, C, FM)	• **Max (AD, FM)**	
Ted 2 (C)	• Dirty Grandpa (C) • Paul Blart: Mall Cop 2 (AC, C, CR) • Spy (AC, C, CR) • The Boss (C)	• Barbershop: The Next Cut (C) • Brooklyn (D, C, R) • Concussion (B, D, SP) • Creed (D, SP) • Fifty Shades of Grey (D, R) • **Focus (C, CR, D)** • Hot Pursuit (AC, C, CR) • Hotel Transylvania 2 (AN, C, FM) • Joy (B, C, D)	• Magic Mike XXL (C, D, MU) • My Big Fat Greek Wedding 2 (C, R) • **Pitch Perfect 2 (C, MU)** • **Ride Along 2 (AC, C)** • Spotlight (B, D, HI, T) • **The Big Short (B, C, D)** • The Intern (C) • The Longest Ride (D, R) • Vacation (AD, C)
The Visit (H, T)	• **Krampus (C, F, H)** • **The Boy Next Door (M, T)** • **The Gift (M, T)** • The Hateful Eight (CR, D, M, W) • **Unfriended (H, M)**	• Southpaw (AC, D, SP, T)	

* Genre Code (AC : Action, AD : Adventure, AN : Animation, B : Biography, C : Comedy, CR : Crime, D : Drama, F : Fantasy, FM : Family, H : Horror, HI : History, M : Mystery, MU : Music, R : Romance, SF : Sci-fi, SP : Sport, T : Thriller, W : Western)

Table 39. Movie recommendations by K-means results using YouTube comments

Source Movie	1st Group, K=11 (Genre)	2nd Group, K=7 (Genre)
Batman v Superman: Dawn of Justice (AC, AD, SF, F)	• Alvin and the Chipmunks: The Road Chip (AN, AD, C, FM) • **Fantastic Four (AC, AD, SF)** • **Pan (AC, AD, FM, F)** • **The Divergent Series: Allegiant (AC, AD, SF, R, M)** • The Hateful Eight (Crime, Drama, Mystery)	• Creed(Drama, Sport) • The Divergent Series: Insurgent (AC, AD, SF, R, M) • Mad Max: Fury Road (AC, AD, SF) • **Terminator Genisys (AC, AD, SF)**
The Peanuts Movie (AN, AD, C, FM)	• Kingsman: The Secret Service(AC, AD, C) • The Good Dinosaur (AN, AD, C, FM) • Tomorrowland (AC, AD, SF, FM)	• **Max (AD, FM)**
Ted 2 (C)	• American Sniper (AC, AD, D) • **Focus (C, CR, D)** • How to Be Single(C, D) • McFarland, USA (B, D, SP) • **Pitch Perfect 2 (C, MU)** • Pixels(AN, AC, C) • The Wedding Ringer(C)	• **Ride Along 2 (AC, C)** • **The Big Short (B, C, D)**
The Visit (H, T)	• **Krampus (C, F, H)** • **The Boy Next Door (M, T)** • **The Gift (M, T)** • **Unfriended (H, M)**	• Everest(AD, BIO, D) • Fifty Shades of Grey (D, RO, T) • Paul Blart Mall Cop 2(AC, C, CR) • The Perfect Guy(D, T)

* Genre Code (AC : Action, AD : Adventure, AN : Animation, B : Biography, C : Comedy, CR : Crime, D : Drama, F : Fantasy, FM : Family, H : Horror, HI : History, M : Mystery, MU : Music, R : Romance, SF : Sci-fi, SP : Sport, T : Thriller, W : Western)

As we mentioned in Section 2.5, the K-NN algorithm was designed to finds relevant objects by a majority vote of its neighbors in order to determine similarity. We also used R-Studio to calculate their similarities to recommend movies [101]. In this experiments, we used 1 to 5 as *k* values (*k* = 1 to *k* = 5) and the lowest *k* value has the highest similarity. Table 40 and 41 show relevant movies based on the results by each *k* value. The movies, "Fantastic Four" and "Krampus," are co-occurred in both the results of movie reviews and YouTube comments (see Table 40 and 41). We assume that these movies have a stronger relation and a higher priority to recommend than the others.

Through these results, we discovered that the movie reviews contain more useful information to recommend movies than YouTube comments because the results of movie reviews seem more relevant compared with their original genres.

Table 40. Movie recommendations using movie reviews (K-NN)

K / Source	K=1	K=2	K=3	K=4	K=5
Batman v Superman: Dawn of Justice	London Has Fallen	Terminator Genisys	The 5th Wave	The Martian	**Fantastic Four**
The Peanuts Movie	Alvin And The Chipmunks: The Road Chip	The SpongeBob Movie: Sponge Out of Water	Paddington	Ted 2	The Duff
Ted 2	Hot Pursuit	Vacation	The SpongeBob Movie: Sponge Out of Water	The Intern	Creed
The Visit	**Krampus**	The Perfect Guy	Bridge of Spies	Taken 3	The Gift

Table 41. Movie recommendations using YouTube comments (K-NN)

K / Source	K=1	K=2	K=3	K=4	K=5
Batman v Superman: Dawn of Justice	Fantastic Four	Ant-man	The Hateful Eight	The Revenant	Pan
The Peanuts Movie	The Good Dinosaur	Tomorrowland	Max	Miracles from Heaven	Kingsman: The Secret Service
Ted 2	American Sniper	Joy	Focus	Concussion	Pitch Perfect 2
The Visit	San Andreas	The Hunger Games: Mockingjay - Part 2	The 5th Wave	Unfriended	Krampus

7. CONCLUSIONS

In this chapter, we summarize the objectives and contributions of this thesis. This study is focused on social media analysis because the social media data contains useful information to understand trends, issues, individuals, human behavior, and identifying influencers. First of all, we analyzed Twitter to discover characteristics of social media in section 4. This study is intended to address these topics to build a better understanding of Twitter usages. Even though there are many important theories and frameworks in media studies, we were not able to single out one specific theoretical framework for this study. Therefore, by using the "active audience concept," and relying on marketing literature, we chose a grounded theory approach of the mass communication field and presented research questions for in-depth understanding of Twitter usage in order to detect any patterns that consumers might show. This interdisciplinary research with the Mass Communication Department helps for better understanding of social media. This analysis has contributed to the study of the Twitter usage pattern by examining message types, URL sources, and devices that people used. As consumers are multi-tasking with several media platforms and their conversations are more fragmented, it becomes more important to have a better understanding of Twitter usage patterns. Also, a recent industry report indicated that people who used Twitter, whether actively tweeting or just following, were 62% more likely to recall the brands which advertised during a TV show than those who did not use Twitter [80]. Therefore, understanding the value of Twitter in marketing and

understanding how consumers use Twitter becomes more important and timely. By analyzing massive data, this study provided a more holistic picture of Twitter usage patterns. As Twitter is using more complicated interfaces through technologies, understanding consumers' media uses and experiences with Twitter will be challenging.

In section 5, we proposed a lexicon building method and a sentiment analysis method using the morphological sentence patterns model. These method aims to observe and summarize people's opinions or emotional states from the social media data. In section 5.1, we proposed a lexicon building method to minimize human-coding efforts. Through the experiments, we found 3 main characteristics. The first characteristic is the length of pattern. When we used "selected patterns" selected by the length through our experiments, the system yield the best results on both processing time and F-score. The second characteristic that more frequently occurred aspects and expressions tend to be more accurate. The third characteristic is that the more frequently co-occurred aspects and expressions tend to be more accurate. We can assume that aspects can be clues for extracting expressions. In addition, we suggested the threshold values, which can adjust what users require whether a better precision or a better recall through our experiments.

To discover how the morphological sentence patterns work across other media, we examined cross-domain analysis using movie reviews, YouTube comments, and tweets. Through the experiments, we discovered that the sentence consists of different

structures in different sources. It implies that people share their opinions and emotions differently depending on the media.

In section 5.2, we also proposed a sentiment analysis method using the morphological sentence pattern model. Through the experiments, we found a characteristic that the meaningful aspects and expressions are mostly adjacent or located within 1 to 2 distances in terms of the part of speech. In addition, we developed a method to solve the partial matching problem and mismatching problem to improve the accuracy. The main advantage of our method is that no human-coded train-set is required to maintain the suggested accuracy. All our proposed methods showed a relatively higher F-score than existing approaches.

In section 6, we propose an extracting method of movie genre similarity for movie recommendation using the morphological sentence pattern model and machine learning algorithms. Our method consists of two main methods, which are “TDF-IDF” and “Genre Score” to discover a similarity of movies with consideration for genres. To experiment our methods, we selected the Top 100 movies listed on the box office released from January 2015 to May 2016 and we collected and analyzed 100 movie reviews and 2,000 YouTube comments for each movie. Through the experiments, we discovered that aspects were more commonly used over genres than expressions. Therefore, we decided to use both aspects and expressions separately as features to calculate the genre score. Also, we verified that movie reviews and YouTube comments contain useful information

to find characteristics of movie genres. Then, we recommended movies based on the results of the K-means and K-NN algorithms.

The main contributions of this research are summarized as follows. First, we proposed a system to handle social media data being generated in real-time using distributed architectures such as Hadoop, and JMS (ActiveMQ). Second, we analyzed Twitter usages patterns with the Mass Communication Department based on the theories of mass communication and the computer science knowledges. Third, we proposed a sentiment analysis method and a lexicon building method for extracting useful information from the social media data. These method show relatively higher accuracy than existing approaches without any human coding efforts. Fourth, we proposed a method for extracting movie genre similarity from social media data using the MSP model. Thus, we verified that social media data contains useful information through all experiment results.

Social media has become an important marketing venue, allowing them to reach a wide range of target audiences efficiently. However, some challenges exists on social media analysis such as a real-time processing and linguistic problems. Thus, we proposed analysis results and methods to understand and discover meaningful information from the data. This research has led us one step closer to facing the challenges on social media analysis.

REFERENCES

[1] Duggan, M. (2015). The demographics of social media users. Mobile Messaging and Social Media 2015, Retrieved from http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/

[2] Twitter Usage Statistics. (2014). Internet Live Stats, Retrieved from http://www.internetlivestats.com/.

[3] M.,Boyd, D. & Ellison, N. B. (2007). Social Network Sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 210-230.

[4] Campbell, C., Pitt, L. F., Parent, M., & Berthon, P. R. (2011). Understanding consumer conversations around ads in a web 2.0 world. *Journal of Advertising*, 40(1), 87- 102.

[5] Chu, S., & 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.

[6] Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons*, 53, 59-68.

[7] Sharma A., & Dey, S. (2012, October). A comparative study of feature selection and machine learning techniques for sentiment analysis. *Proceedings of the 2012 ACM Research in Applied Computation Symposium* (RACS), 1-7.

[8] Goncalves, P., Araújo, M., Benevenuto, F., & Cha, M. (2013, October). Comparing and combining sentiment analysis methods. *Proceedings of the 1st ACM conference on online social networks*, 27-38.

[9] Melville, P., Gryc, W., & Lawrence, R. D. (2009, June). Sentiment analysis of blogs by combining lexical knowledge with text classification. *ACM international conference on Knowledge discovery and data mining (SIGKDD'09)*, 1275-1284.

[10] Han, Y., Kim, Y., & Jang, I. (2016), A Method for Extracting Lexicon for Sentiment Analysis based on Morphological Sentence Patterns. *Studies in Computational Intelligence* (SCI): *Software Engineering Research, Management and Applications (SERA)*, Springer, 654, 85-101.

[11] Han, Y., Kim, Y., & Song, J. (2017). A Cross-Domain Analysis using Morphological Sentence Pattern Approach for Extracting Aspect-based Lexicon. *International Journal of Multimedia and Ubiquitous Engineering (IJMUE)*, In-press.

[12] Ferguson, D. A., & Perse, M. (2000). The World Wide Web as a functional alternative to television. *Journal of Broadcasting and Electronic Media*, 44(2), 155-174.

[13] Papacharissi, Z., & Rubin, A. M. (2000). Predictors of Internet use. *Journal of Broadcasting & Electronic Media*, 44(2), 175-196.

[14] Miller, R., & Washington, K. (2013). Consumer use of media & the Internet. Entertainment. *Media & Advertising Market Research Handbook*, 13, 21-28.

[15] Advertising Age. (2016). Marketing Fact Pack, Retrieved from http://adage.com/d/resources/resources/whitepaper/2016-edition-marketing-fact-pack

[16] Baran, S. J., & Davis, D. K. (2011). Mass Communication Theory: Foundations, Ferment, and Future. Wadsworth, Cengage Learning.

[17] Ko, H., Cho, C., & Roberts, M. S. (2005). Internet uses and gratifications: a structural equation model of interactive advertising. *Journal of Advertising*, 24 (2), 57-70.

[18] Hollenbaugh, E. E. (2010). Personal journal bloggers: Profiles of disclosiveness. *Computers in Human Behavior*, 26(6), 1657-1666.

[19] Chen, G. M. (2011). Tweet this: A uses and gratifications perspective on how active Twitter use gratifies a need to connect with others. *Computers in Human Behavior*, 27(2), 755-762.

[20] Muntinga, D. G, Moorman, M., & Smit, E. G. (2011). Introducing COBRAs, *International Journal of Advertising*, 30(1), 13-46.

[21] Stafford, T. F., Stafford, M. R., & Schkade, L. (2004). Determining uses and gratifications for the Internet. *Decision Sciences*, 35(2), 259-288.

[22] Shao, G. (2009). Understanding the appeal of user-generated media: a uses and gratification perspective. *Internet Research*, 19(1), 7-25.

[23] Schultz, E. J. (2016, January 4). See the 'crash the Super Bowl' finalists' ads. *Advertising Age*, Retrieved from http://adage.com/article/special-report-super-bowl/crash-super-bowl-finalist-ads/301978/.

[24] Barnes, N. G., & Mattson, E. (2009, November 12). The Fortune 500 and blogging: slow and steady and farther along than expected. Retrieved from http://www.umassd.edu-/cmr/studiesresearch/fortune500.pdf

[25] Stelzner, M. (2013). The 2013 social media marketing industry report. Social Media Examiner. Retrieved from http://www.socialmediaexaminer.com/SocialMedia MarketingIndustryReport2013.pdf

[26] Cone Business. (2008). 2008 business in social media study. Retrieved from http://www.coneinc.com/news/request.php?id=1183

[27] eMarketer. (2015 April). Social Network Ad Spending to Hit $23.68 Billion Worldwide in 2015. Retrieved from http://www.emarketer.com/Article/Social-Network-Ad-Spending-Hit-2368-Billion-Worldwide-2015/1012357#sthash.lBpD Flob.dpuf.

[28] Homburg, C., Ehm, L., & Artz, M. (2015). Measuring and Managing Consumer Sentiment in an Online Community Environment. *Journal of Marketing Research*, Vol. LII, 629-641.

[29] Kumar, A., Bezawada, R., Rishika, R., Janakiraman, R., & Kannan, P.K. (2016, January). From Social to Sale: The Effects of Firm-Generated Content in Social

Media on Customer Behavior. *Journal of Marketing,* 80, 7-257 DOI: 10.1509/jm.14.0249

[30] Pehlivan, E., Sarican, F., & Berthon, P. (2011). Mining messages: Exploring consumer response to consumer- vs. firm-generated ads. *Journal of Consumer Behaviour*, 10, 313-321, Published online in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/cb.379

[31] Joinson, A.N. (2008, April 05-10). 'Looking at', 'Looking up' or 'Keeping up with' people? Motives and Uses of Facebook. *In Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems (CHI '08),* 1027- 1036.

[32] Wohn, D. Y., & Na, E. (2011). Tweeting about TV: Sharing television viewing experiences via Social media message streams, First Monday, 16(3), 1-14.

[33] Buschow, C., Schneider B., & Ueberheide, S. 2014. Tweeting television: Exploring communication activities on Twitter while watching TV. Communications-The European Journal of Communication Research 39, 2, 129-149

[34] Schirra, S., Sun, H., & Bentley, F. (2014, April 26 - May 01). Together Alone: Motivations for Live-Tweeting a Television Series. *In Proceedings of the 32th Annual SIGCHI Conference on Human Factors in Computing* Systems *(CHI'14),* 2441-2450.

[35] Doughty, M., Rowland, D., & Lawson S. (2011, June 29 - July 01). Co-Viewing Live TV with Digital Backchannel Streams. *In Proceedings of the 9th international interactive conference on Interactive television (EuroITV'11),* 141-144.

[36] Atkinson, C. (2010). The Backchannel: How Audiences Are Using Twitter and Social Media and Changing Presentations Forever. *New Riders*.

[37] Bruns, A., & Stieglitz, S. (2013). Towards More Systematic Twitter Analysis: Metrics for Tweeting Activities. *International Journal of Social Research Methodology*, 16, 2, 91-108.

[38] Ives N. (2013). Super Bowl Ratings Decline. *Advertising Age*, Retrieved from http://adage.com/article/special-report-super-bowl/super-bowl-ratings-decline-year/239592/

[39] Steinberg B. (2012). CBS Claims Record Super Bowl Ratings in Early Tally. *Advertising Age*, Retrieved from http://adage.com/article/special-report-super-bowl/cbs-claims-super-bowl-ratings-victory-early-tally/239578/

[40] Manning, C. D., Surdeanu M., Bauer, J., Finkel, Jenny., Bethard, S. J., & McClosky, David. (2014, June). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60.

[41] Guerra, P. H., Veloso, A., Meira W., & Almeida, V. (2011, August). From Bias to Opinion: A Transfer-Learning Apporach to Real-Time Sentiment Analysis.

Proceddings of the 17th ACM international conference on Knowledge discovery and data mining (SIGKDD'11), 150-158, ISBN:978-7-4503-0813-7

[42] Kucuktunc, O., Cambazoglu, B. B., Weber I., & Ferhatosmanoglu, H. (2012, February). A Larege Scale Sentiment Analysis for Yahoo! Answers. *Proceedings of the fifth ACM international conference on Web search and data mining*, 633-642, ISBN: 978-1-4503-0747-5

[43] Speriosu M., Sudan, N., Upadhyay, S., & Baldridge, J. (2011, July). Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph. *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, 53-64, ISBN: 978-1-937284-13-8

[44] O'Connor, B., Balasubramanyan, R., Routledge, B. R., & Smith, N. A. (2010). From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 122-129.

[45] Winson, T., Hoffman, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff E., & Patwardhan, S. (2005, October). OpinionFinder:A System for Subjectivity Analysis. *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, 34-35

[46] Lee, H., Han, Y., & Kim, K. (2014). Sentiment Analysis on Online Social Network Using Probability Model. *In Proceedings of the Sixth International Conference on Advances in Future Internet (AFIN)*, 14-19.

[47] Bross, J., & Ehrig, H. (2013, October 27 - November 01). Automatic Construction of Domain and Aspect Specific Sentiment Lexicons for Customer Review Mining. *Proceedings of the 22nd ACM international conference on Information & Knowledge Management (CIKM'13)*, 1077-1086, ISBN: 978-1-4503-2263-8

[48] Thet, T. T., Na, J., & Khoo, C. S. G. (2010) Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6), 823-848.

[49] Wogenstein, F., Drescher, J., Reinel, D., Rill, S., & Scheidt, J. (2013, August 11). Evaluation of an Algorithm for Aspect-Based Opinion Mining Using a Lexicon-Based Approach. *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM '13),* 5, ISBN : 978-1-4503-2332-1

[50] Kaji, N., & Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. *In Proceedings of EMNLP-CoNLL*, 1075–1083.

[51] Xu, J., Xu, R., Zheng, Y., Lu, Q., Wong, K., & Wang, X. (2013). Chinese emotion lexicon developing via multi-lingual lexical resources integration. *In Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing*, 2, 174–182.

[52] Xu, L., Lin, H., Pan, Y., Ren, H., & Chen, J. (2008). Constructing the affective lexicon ontology. *Journal of the China Society for Scientific and Technical Information*, 27(2), 180-185.

[53] Zhang, Z., & Singh, M. P. (2014). Renew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis. *In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 542–551.

[54] Tai, Y., & Kao, H. (2013). Automatic Domain-Specific Sentiment Lexicon Generation with Label Propagation. *Proceedings of International Conference on Information Integration and Web-based Applications & Services (IIWAS '13)*, 53, ISBN: 978-1-4503-2113-6.

[55] Jones, S. K. (1972). A statistical interpretation of term specificity and its application in retrieval, *Journal of Documentation*, 28, 1, 11–21.

[56] Manning, C. D., Raghavan, P., & Schutze, H. (2008). Scoring, term weighting, and the vector space model. *Introduction to Information Retrieval*, 100, ISBN=9780511809071, DOI=10.1017/CBO9780511809071.007

[57] Tang, D., Wei, F., Qin, B., Zhou, M., & Liu, T. (2016, August). Building Large-Scale Twitter-Specific Sentiment Lexicon. *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, 172–182.

[58] Pitsilis, G., Zhang, X., & Wang, W. (2011). Clustering Recommenders in Collaborative Filtering Using Explicit Trust Information. *IFIP Advances in Information and Communication Technology (IFIPAICT*), 358, 82-97.

[59] Forgy, E.W. (1967). Cluster analysis of multivariate data: efficiency versus interpretability of classifications, *Biometrics 21, In Biometric Society Meeting,* 768-769.

[60] MacQueen, J. B. (1967). Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, *University of California Press*, 281–297. DOI=10.1.1.308.8619

[61] Lloyd, S. P. (1982). Least squares quantization in pcm, *IEEE Transactions on Information Theory*, 28, 2, 129-136, DOI= 10.1109/TIT.1982.1056489

[62] Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in Strategic Management Research: An analysis and critique, *Strategic Management Journal*, 17, 6, 441–458, DOI=10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

[63] Wahl, H., Winiwarter, W., & Quirchmayr, G. (2010, November). Natural Language Processing Technologies for Developing a Language Learning Environment, *Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services,* 381-388, ISBN: 978-1-4503-0421-4

[64] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician,* 46 (3): 175–185. DOI=10.1080/00031305 .1992.10475879.

[65] Wen, Z. (2008). Recommendation System Based on Collaborative Filtering, *Stanford CS229 Projects*.

[66] Ricci, F., Rokach, L., & Shapira, B. (2001) Introduction to Recommender Systems Handbook, *Recommender Systems Handbook (Springer),* 1-35

[67] Terveen, L., & Hill, W. (2001). Beyond Recomendador Systems: Heling People Help Each Other. *Human-Computer Interaction in the New Millennium (Addison-Wesley),* 22, 487-509

[68] Han, Y., Lee, H, & Kim, Y. (2015, October). A Real-time Knowledge Extracting System from Social Big Data using Distributed Architecture. *Proceedings of the 2015 Research in Adaptive and Convergent Systems (RACS'15)*, 74-79.

[69] King, I., Li, J., & Chan, K. Tong. (2009). A brief survey of computational approaches in social computing. *Proceedings of the 2009 international joint conference on Neural Networks*, 2699-2706.

[70] Byun, C., Lee, H., & Kim, Y. (2012). Automated Twitter data collecting tool for data mining in social network. *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, 196-204.

[71] Lee, H., Han, Y., Kim, K. K., & Kim, Y. (2014). Sports and Social Media: Twitter Usage Patterns during the 2013 Super Bowl Broadcast. *Proceedings of the 4th International Conference on Communication, Media, Technology and Design (ICCMTD'14)*, 250-259

[72] O'Connell, M. (2014, February 3). TV ratings: Super Bowl XLVIII is most watched in history with 112.2 million viewers. *The Hollywood Reporter*, Retrieved from http://www.hollywoodreporter.com/live-feed/tv-ratings-super-bowl-xlviii-676651

[73] Dumenco, S. (2011, December 14). And the No. 1 Social-TV Hit of the Year Is… (Hint: Not the Oscars or Super Bowl). *Advertising Age*, Retrieved from http://adage.com/article/trending-topics/1-social-tv-hit-year/231574/.

[74] Eversley, M. (2013, February 4). Super engagement record for Super Bowl. USA Today. Retrieved from http://www.usatoday.com/story/money/business/2013/02/04/super-engagement-social-media/1890351/

[75] Learnonth, M. (Feb. 25, 2013). When did Twitter grow up? Answer: Now. Marketing has become an in-the moment game, thanks to one company, *Advertising Age*, Retrieved from http://adage.com/article/digital/twitter- grow/239992/

[76] Adobe Digital Index. (2014, January 31). Digital & TV 'double coverage' key to marketing success at the Super Bowl social media is off the sidelines this year. *Advertising Age*, Retrieved from http://adage.com.proxy-tu.researchport.umd.edu/article/adobe/adobe-digital-tv-key-marketing-success-super-bowl/291429/

[77] Kwak, H., Lee, C., Park, H., & Moon, S. (2010, April). What is Twitter, a social network or a news media? *Proceedings of the 19th international conference on World Wide Web (WWW'10),* 591-600.

[78] Larsson, A. O. (2013). Twitting the viewer –Use of Twitter in a talk show context. *Journal of Broadcasting & Electronic Media*, 57 (2), 135-152.

[79] CBS claims record Super Bowl ratings in early tally: Early figures do not include half hour during blackout. (February 04, 2013). *Advertising Age*. Retrieved from www.adage.com.

[80] Warc. (2016, March 21). TV Tweeters have higher ad recall. *Warc.com*. Retrieved from http://www.warc.com/LatestNews/News/EmailNews.news?ID=36426&Origin=WARCNewsEmail&CID=N36426&PUB=Warc_News&utm_source=WarcNews&utm_medium=email&utm_campaign=WarcNews20160321.

[81] Tomovic, A., Janicic, P., & Keselj, V. (2006) n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Journal of computer methods and programs in biomedicine*, 81, 137-153.

[82] Hu, M., & Liu, B. (2004, August). Mining and summarizing customer reviews. *ACM international conference on Knowledge discovery and data mining (SIGKDD'04)*, 168-177.

[83] Mohammad, S.M., & Turney, P. D. (2012). Crowdsourcing a word–emotion association lexicon. *Computliuational Intelligence*, 29(3), 436-465.

[84] Tang, D., Wei, F., Qin, B., Zhou, M., & Liu, T. (2014, August). Building Large-Scale Twitter-Specific Sentiment Lexicon. *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, 172–182.

[85] Song, J. J. (2012) Word order, *Cambridge University Press*, ISBN:978-0-521-87214-0

[86] Gonçalves, P., Benevenuto, F., & Cha, M. (2013). Panas-t: A pychometric scale for measuring sentiments on twitter. *Social and Information Networks-Physics and Society (arXiv:1308.1857)*, 14

[87] Hogenboom, A., Bal, D., Frasincar, F., Bal, M., de Jong, F., & Kaymak, U. (2013). Exploiting emoticons in sentiment analysis. *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC'13),* 703-710

[88] Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012). A system for real-time twitter sentiment analysis of 2012 U.S. presidential election cycle. *Proceedings of the ACL 2012 System Demonstrations (ACL'12)*, 115–120.

[89] Cambria, E., Speer, R., Havasi, C., & Hussain, A. (2010) Senticnet: A publicly available semantic resource for opinion mining. *In: AAAI fall symposium: commonsense knowledge*, 10, 2.

[90] Esuli, A., & Sebastiani, F. (2006). Sentwordnet: A publicly available lexical resource for opinion mining. *In International Conference on Language Resources and Evaluation (LREC'06)*, 417–422.

[91] Thelwall, M. (2016). The Heart and Soul of the Web? Sentiment Strength Detection in the Social Web with SentiStrength. *Cyberemotions (Part of the series Understanding Complex Systems)*, 119-134.

[92] Dodds, P. S., & Danforth, C. M. (2009). Measuring the happiness of large-scale written expression: songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4), 441–456.

[93] Tausczik, Y. R., and Pennebaker, J. W. (2010). The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1), 24–54.

[94] Liu A., Zhang Y., & Li. J. (2009, October). Personalized movie recommendation. *In Proceedings of the 17th ACM, International Conference on Multimedia (MM'09),* 845–848. DOI=http://doi.acm.org/10.1145/1631272.1631429

[95] Mukherjee R., Sajja N., & Sen S. (2003). A Movie Recommendation System-An Application of Voting Theory in User Modeling. *User Modeling and User-Adapted Interaction*, 13, 1, 5–33. DOI=10.1023/A:1024022819690

[96] Orellana-Rodriguez C., Diaz-Aviles E., & Nejdl W. (2015). Mining Affective Context in Short Films for Emotion-Aware Recommendation. *Proceedings of the 26th ACM Conference on Hypertext & Social Media (HT '15),* 185-194, DOI=http://doi.acm.org/10.1145/2700171.2791042

[97] Diao, Q., Qiu, M., Wu, C., Smola, A., Jiang, J., & Wang, C. (2014, August). Jointly Modeling Aspects, Ratings and Sentiments for Movie Recommendation (JMARS). *ACM international conference on Knowledge discovery and data mining (SIGKDD'14),* 193-202. DOI=http://doi.acm.org/10.1145/2700171.2791042

[98] Sebastiani, F. (2002). Machine learning in automated text categorization, *ACM Computing Surveys*, 34, 1, 1–47, DOI=10.1145/505282.505283

[99] Brezeale, D., & Cook, D. J. (2008). Automatic Video Classification: A Survey of the Literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38, 3, 416-430, DOI=10.1109/TSMCC.2008.91917

[100] Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S. *Statistics and Computing (Springer, 4th Ed.)*, DOI= 10.1007/978-0-387-21706-2

[101] R Core Team (2016). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Retrieved from https://www.R-project.org/.

[102] Maechler M., Rousseeuw P., Struyf A., Hubert M., & Hornik K. (2013). Cluster: Cluster Analysis Basics and Extensions, R package version 1.14.4.

[103] Hadoop, http://hadoop.apache.org

[104] X. Liu, N. Iftikhar and X. Xie (2014). Survey of real-time processing systems for big data, Proceedings of the 18th International Database Engineering & Applications Symposium, 356-361, ISBN: 978-1-4503-2627-8

[105] Mark, R., Richard, M., and David, A. C. (2009), The Advantages of Messaging, *in Java Message Service, 2nd ed., Ed. California: O'Reilly Media*, 3-5, ISBN:978-0-596-52204-9

[106] Michael, M., Robert, H., Christian, Z., & Sebastian, G., "Throughput Performance of Popular JMS Servers," Proceedings of the joint international conference on Measurement and modeling of computer systems, June 2006, pp 367-368, ISBN:1-59593-319-0

[107] Mirco, M., Cecilia, M., Stephen, H. (2004), Adapting asynchronous messaging middleware to ad hoc networking, Proceedings of the 2nd workshop on Middleware for pervasive and ad-hoc computing, 121–126,ISBN:1-58113-951-9

[108] M. Song, M. Kim and Y. Jeong, "Anlyzing the Political Landscape of 2012 Korean Presidential Election in Twitter", Intelligent Systems, IEEE, vol 29, Issue 2, pp.18-26, March 2014.

[109] M. Boanjak and E. Oliveira. "TwitterEcho - A distributed focused crawler to support open research with twitter data", International conference companion on World Wide Web, April 2012, pp. 1233–1239, ISBN: 978-1-4503-1230-1

[110] C. Byun, H. Lee, Y. Kim, and K. K. Kim. "Twitter data collecting tool with rule-based filtering and analysis module", International Journal of Web Information Systems, Vol 9, Issue 3, pp. 184-203, 2013

[111] A. Black, C. Mascaro, M. Gallagher, and S. P. Goggins. “Twitter Zombie: Architecture for capturing, socially transforming and analyzing the Twittersphere”, International conference on Supporting group work, October 2012, pp. 229–238. ISBN:978-1-4503-1486-2

[112] Y. Stavrakas and V. Plachouras, “A platform for supporting data analytics on twitter challenges and objectives” Intl. Workshop on Knowledge Extraction & Consolidation from Social Media, (Ict 270239), 2013.

[113] D. Preotiuc-Pietro, S. Samangooei, and T. Cohn, “Trendminer : An architecture for real time analysis of social media text”, Workshop on RealTime Analysis and Mining of Social Streams, 2012,pp. 4–7.

[114] K. Bontcheva and L. Derczynski, “TwitIE: an opensource information extraction pipeline for microblog text”, International Conference on Recent Advances in Natural Language Processing, 2013.

[115] J. Yin, S. Karimi, B. Robinson, and M. Cameron “ESA: emergency situation awareness via microbloggers” Proceedings of the 21st ACM international conference on Information and knowledge management, October 2012, pp. 2701–2703.

[116] T. Baldwin, P. Cook, and B. Han “A support platform for event detection using social intelligence,” Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, April 2012, pp 69–72.

[117] L. Bing, “Sentiment Analysis and Subjectivity,” An chapter in Handbook of Natural Language Processing, Second Edition, 2010.

CURRICULUM VITA

Youngsub Han

[redacted]

Department of Computer and Information Sciences

Towson University

Education

Ph. D. in Computer Science
Towson University, Maryland, US (May 2017)

M. S. in Computer Science
Towson University, Maryland, US (May 2009)

B. Sc. in Computer Science
Shinhan University, Gyeonggi-do, South Korea (February 2006)

Publications

Journals

1. **Han, Y.,** & Kim, Y. (2017), An Extracting Method of Movie Genre Similarity using Aspect-Based Approach in Social Media, *SIGAPP Applied Computing Review Journal (ACR),* 17, 2 (In press)
2. **Han, Y.,** Hong, B., Lee, H. & Kim, K. (2017), How do we tweet? The comparative analysis of Twitter usage by message types, devices, and sources, *The Journal of Social Media in Society (JSMS),* 6, 1 (In press)
3. **Han, Y.,** Kim, Y., & Jang, I. (2016), A Method for Extracting Lexicon for Sentiment Analysis based on Morphological Sentence Patterns. *Studies in Computational Intelligence (SCI): Software Engineering Research, Management and Applications, Springer,* 654, 85-101.

4. Choi, J., **Han, Y.,** & Kim, Y. (2016), A Research for Finding Relationship between Mass Media and Social Media based on Agenda Setting Theory, *Studies in Computational Intelligence (SCI): Software Engineering Research, Management and Applications, Springer,* 654, 103-113.

Proceedings

---

1. **Han, Y.,** & Kim, K (2017) Sentiment Analysis on Social Media Using Morphological Sentence Pattern Model, Software Engineering Research, Management and Applications (SERA 2017), (Accepted)
2. **Han, Y.,** Hong, B., & Kim, K. (2017), A Study of the Viewers' Social TV Behaviors during a Sporting Event, Social Media & Society (Accepted)
3. **Han, Y.,** & Kim, Y. (2016), A Method of Discovering Genre Similarity using Aspect Based Approach, Proceedings of the 2015 Research in Adaptive and Convergent Systems (RACS'16), 29-34.
4. **Han, Y.,** Kim, & Y. Song, J. (2016), Building Sentiment Lexicon for Social Media Analysis using Morphological Sentence Pattern Model, The 5th International Conference on Information Technology and Computer Science (ITCS 2016), 103-106
5. **Han, Y.,** Lee, H, & Kim, Y. (2015, October). A Real-time Knowledge Extracting System from Social Big Data using Distributed Architecture. Proceedings of the 2015 Research in Adaptive and Convergent Systems (RACS'15), 74-79.
6. Hong, B., **Han, Y.,** & Kim, Y. (2015) A Semi-supervised Tweet Classification Method Using News Articles, Proceedings of the 2015 Research in Adaptive and Convergent Systems (RACS'15), 62-67.
7. Lee, H., **Han, Y.,** & Kim, K. (2014). Sentiment Analysis on Online Social Network Using Probability Model. In Proceedings of the Sixth International Conference on Advances in Future Internet (AFIN), 14-19.

8. Lee, H., **Han, Y.,** Kim, K. K., & Kim, Y. (2014). Sports and Social Media: Twitter Usage Patterns during the 2013 Super Bowl Broadcast. Proceedings of the 4th International Conference on Communication, Media, Technology and Design (ICCMTD'14), 250-259

Conference Presentations

1. How Do We Tweet? The Comparative Analysis of Twitter Usage Patterns by Message Types, Sources, and Devices, Presented at the Association for Education in Journalism and Mass Communication (AEJMC) Southeast Colloquium, Louisiana State University, Baton Rouge, LA, US, March 3-5, 2016
2. Twitter Usage Patterns by User Relationships in Social Media, Presented at the 10th Interdisciplinary Social Sciences Conference, Split, Croatia, June 9-12, 2015