TOWSON UNIVERSITY

COLLEGE OF GRADUATE STUDIES AND RESEARCH

TEXT CLASSFICATION AND SENTIMENT ANALYSIS IN SOCIAL NETWORKS USING A PROBABILITY MODEL

By

Hyeoncheol Lee

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 2015

TOWSON UNIVERSITY
OFFICE OF GRADUATE STUDIES

DISSERTATION APPROVAL PAGE

This is to certify that the dissertation prepared by Hyeoncheol Lee , entitled "Text Classification and Sentiment Analysis in Social Networks Using a Probability Model " has been approved by the thesis committee as satisfactorily completing the dissertation requirements for the degree of Doctor of Science in Information Technology.

	4-22-2015
Chairperson, Dissertation Committee, Dr. Yanggon Kim	Date
	4-22-2015
Committee Member, Dr. Sungchul Hong	Date
	4-22-2015
Committee Member, Dr. Siddharth Kaza	Date
	4-22-2015
Committee Member, Dr. Michael McGuire	Date
Janet V DeLany May 14, 20	
Dean of Graduate Studies	Date

Acknowledgement

First of all, I would like to gratefully thank my advisor, Dr. Yanggon Kim, for his excellent 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. Also, 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, Youngsub Han and Beomseok Hong. My completion of this dissertation could not have been accomplished without support of them.

I also would like to express my gratitude to IsciLab members Dr. Hyeong-In Choi, Dr. Hyeong-Ah Choi, Dr. Wonpyo Hong, Dr. Changhyun Byun, Dr. Changmin Lee and Dr. Jekuk Yun. I am so deeply grateful for their help, professionalism, encouragement and financial support throughout the challenging projects.

Finally, and most importantly, I would like express deepest gratitude to my parents, Manseok Lee and Nomi Jung, grandmother, Yeong-Gui Shin, sisters, Jungmi, Malhee, Jeongsook and Misoo, and their husbands, Inho Park, Dongwook Kim, Seungdae Shin and Seokyong Yoon. They were always supporting me and encouraging me with their best wishes.

This dissertation is dedicated to my family

ABSTRACT

TEXT CLASSFICATION AND SENTIMENT ANALYSIS IN SOCIAL NETWORKS USING A PROBABILITY MODEL

Hyeoncheol Lee

In recent years, diverse social networks, such as Facebook, YouTube, and Twitter, have rapidly grown in size and influence, and a huge amount of data is being generated from the social networks in real time. Demands for data mining on social networks have been dramatically increasing, since analyzing the data can yield insights and understanding to real world phenomenon. However, there are a lot of challenges and difficulties with data collection, management, and analysis because of the features of the social networks' data: large, noisy, and dynamic. Therefore, this study will address the overall problems with data mining in social networks and improve existing data mining techniques. We propose an integrated data collection, management and analysis system. Furthermore, we propose specific analysis methods, such as topic classification, sentiment analysis, and seed selection algorithm to analyze social networks' data.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

In recent years, diverse social networks, such as Facebook, YouTube, and Twitter, have rapidly grown in size and influence, and they have changed the way people communicate with each other. With the increasing popularity of online social network sites, a huge amount of data is being generated from them in real time. Analyzing the data in social media can yield interesting perspectives to understanding human behavior, detecting hot topics, identifying influential people, and/or discovering a group or community[57][58].

Although demands for social network data analysis have been dramatically increasing, methodologies of data collection, management and analysis are not yet maturely established because of several reasons. First of all, data volume generated from social networks has exponentially increased in the last decade. For this reason, an efficient data crawler that collects a topic’s relevant data became an important task in the data collection process. Not to mention that enhancing the existing data store and process of data is essential to handle and analyze a huge amount of data ranging from a few terabytes to multiple petabytes. Since a lot of online social networks and web sites are available, collecting data in multiple social networks also has become an essential task. For these reasons, we propose an integrated data collection and management system that can efficiently crawl and handle texts documents from multiple social network sources.

The topic model uncovers abstract topics within texts documents [28][29][30], which is an important task in text analysis in social networks. However, the characteristics of text documents in social networks are different from traditional text documents, which makes the topic classification process challenging. The length of the documents is short in that it ranges from several words to a few sentences. Terminologies in the text documents are rapidly changing as trends of social networks change. Furthermore, most of the text documents are unlabeled and unstructured, which also makes the topic model in online social networks problematic. Thus, we propose a topic classification algorithm, based on several machine learning algorithms, regarding the features of text documents in social networks.

Several user-generated text documents in social networks contain users' emotional state and mood about topics, such as events, products, and services. These documents can be used to understand and predict real world phenomenon. Sentiment analysis is used to extract people's opinion and knowledge from text documents [17][18]. Recently, demands for an automated sentiment analysis tool for text documents generated from the web have dramatically increased and the literature on this topic has been growing. However, the nature of text documents in social networks makes sentiment analysis difficult and the accuracy of previous automatic sentiment analysis approaches remains around 80%. This should be further improved for more accurate analysis. In addition, some existing approaches require additional information, such as users' tendencies or relationships, which are not always available on online social networks. For these reasons, we propose a sentiment analysis algorithm that guarantees higher accuracy than existing approaches and can be used broadly in social network sites without requiring additional information.

Twitter is a microblogging service and one of the most popular online social network services at present. Twitter allows users to write a message of up to 140 characters, which is known as a tweet. One of main characteristics that differentiate Twitter from other online social network services is the following relationships. If user A follows user B, user A is a follower and receives all tweets user B writes. User B does not have to follow back user A. Once a user writes a tweet, the followers of the user can view the tweet. People write tweets about their life or social issues and the tweets are spread out on Twitter by the following relationships. Twitter also provides an application programming interface (API) to allow developers access a variety of Twitter’s data. Numerous researchers have studied gathering and analyzing the Twitter data to detect current issues such as earthquakes [59] and influenza [60] or to recommend appropriate tags to users [61].

Many researchers have developed their own tweet crawlers [2][12][53][55]. Recently, a change in Twitter's terms of service prohibits publicly sharing crawled data for any purposes, including academic research [2]. Moreover, there is still an additional restriction imposed by Twitter; access to Twitter APIs is limited to a relatively reduced number of queries. A general-purpose tweet crawler that would collect all tweets around the world is unlikely to be legally feasible to implement. On the contrary, a subject-oriented tweet crawler, which collects tweets that are relevant to given search terms, is attractive to researchers.

Among the many considerations to gather the tweets relevant to a specific topic, the selection of seed nodes, which are used for the starting point of the data gathering process, is the most important issue to investigate. In order to effectively collect topic-related data from Twitter, this paper proposes an algorithm to select suitable seed nodes, which can greatly improve the efficiency of the crawling process. The algorithm considers user influences and activities to find the best seed nodes dynamically.

Overall, we propose an integrated data collection, analysis, and management system, as well as specific data analysis and processing techniques in a broad spectrum. This study is structured as follows. Section 2 contains a description of work related to social network and text analysis methods. Section 3 explains design and implementation details of the data collection and management system. Section 4 describes the text classification algorithm that categorizes given documents into specific topics. We propose a sentiment analysis algorithm that analyzes the polarity of documents toward a topic in Section 5. Section 6 proposes a seed selection algorithm for efficient data collection. In Section 7, several data analysis case studies are explained. Section 8 concludes this study with summarization, main contributions, limitations, and future research topics.

2. Literature Review

2.1. Data Collection and Management

These days, social data is being generated, produced, and exchanged enormously [1]. Demands for extracting meaningful information from social data has been dramatically increased with the huge amount of data generated from social networks. A lot of researchers have developed data collection and management systems for online social network analysis, which are summarized in Table 1. Most previous research encompass potential problems with data processing, management, and analysis. First, the data storage of the previous approaches is 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. Second, they do not support distributed processing, which may increase processing time. Last, 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.

Table 1. The Comparison of Data Collection and Management Tool

	Source Channel	Data Store	Distributed Processing
Song et al [1].	Twitter	Relational, Key-value pairs	No
TwitterEcho [2]	Twitter	Not given	No
Byun et al [3].	Twitter	Relational	No
Twitter Zombie [4]	Twitter	Relational	No
TwitHoard [5]	Twitter	Graph DB	No
TrendMiner [6]	Twitter	Key-value pairs	No
TwitIE [7]	Twitter	Not given	No
ESA [8]	Twitter	Not given	No
Baldwin et al [9].	Twitter	Flat files	No

Apache Hadoop is an open source software that allows us to process and manage data in scalable and distributed manners [10]. The main characteristic that separates Apache Hadoop from other database management systems is that distributed processing of large data sets is allowed across clusters of computers with simple programming models. It also supports job scheduling and multiple cluster resource management. The distributed computing techniques provide high scalable processing capabilities that reduces processing time for big data [11]. Hadoop uses the Hadoop Distributed File System (HDFS) that splits files into large blocks and distributes blocks into nodes in clusters. It runs a range of clusters in commodity machine. MapReduce is a distributed data processing model and execution environment that processes data using parallel processing in the nodes. Hadoop is widely used in industry to process and manage big data. It can be considered a candidate for social data analysis since it requires handling and analyzing a vast amount of data.

2.2. Natural Language Processing

Natural Language Processing (NLP) is the process of analyzing and understanding language that humans use with computer algorithms [12][13][14][15]. Since text documents written by humans are generated from online social networks, NLP is widely used in social network analysis. Approaches to NLP vary, reaching from word and token-based analysis to hierarchical or logical representation analysis.

The Stanford Natural Language Processing Group has developed Natural Language Processing Software that provides us with statistical, deep-learning, and rule-based NLP tools [16]. The software is widely used in industry, academia, and government. Key software modules are summarized in Table 2.

Table 2. List of Stanford Natural Language Processing Software Modules and Features

Module Names	Features
Stanford CoreNLP	An integrated suite of natural language processing tools
Stanford Parser	Probabilistic natural language parsers
Stanford POS Tagger	A maximum-entropy (CMM) part-of-speech (POS) tagger
Stanformd Named Entity Recognizer	A Conditional Random Field sequence model
Stanford Word Segmenter	A CRF-based word segmenter
Stanford Classifier	A machine learning classifier. Provides a softmax classifier, Naïve Bayes, and other options.
Tregex, Tsurgeon, and Semgrex	Tools for matching patterns in linguistic trees
Phrasal	Phrase-based machine translation system
Stanford EnglishTokenizer	Tokenizer for English text
Stanford Token Regex	Matching regular expressions over tokens
Stanford Temporal Tagger(SUTime)	A rule-based temporal tagger for English text
Stanford Pattern-based information Extraction and Diagnostics(SPIED)	A boostrapped pattern-based entity extraction system.
Stanford Relation Extractor	A tool for extracting relations between entities

2.3. Sentiment Analysis

Several user-generated text documents contain users' emotional states and moods about topics, such as events, products, and services. Sentiment analysis is extracting the users' opinion and knowledge from the text documents [17][18]. Most approaches focus on identifying whether a text document expresses a positive or negative opinion about a topic [18][19]. The high volume of such data has called for automated tools that assign positive or negative for much easier and quicker analysis. Recently, the literature on this topic has been growing.

There are two main approaches to extracting sentiment from text documents. The first approach is lexicon-based sentiment analysis which is found on pattern matching with pre-built lexicon. Many researches tried to extract sentiment or opinion from text documents using this approach [17][18][19][20]. O'Connor et al. [20] analyzed political opinion using a sentiment analysis algorithm. They collected text documents related to political opinion in Twitter from 2008 to 2009. Also, they built a sentiment lexicon where each word was categorized as either positive or negative keywords based on OpinionFinder [21]. The number of positive and negative keywords was counted for every text document. A text document is defined as positive if it contains any positive word, and negative if it contains any negative word. As a result, the ratio of positive documents versus negative documents was compared with survey results and it showed data correlation between the results of sentiment analysis and survey is as high as 80%. The results indicate that the method can be used as a supplement for a traditional survey. However, the lexicon based approach has weaknesses in that a text document including positive keywords does not necessarily yield positive opinion. For instance, the word *like* is categorized as a positive word in the lexicon, meaning if a text document includes the word *like,* it is categorized as a positive document. Nevertheless, if the message includes the word *do not* right before *like*, the actual opinion of the document should be categorized as negative. In this sense, such lexicon-based approaches should be improved regarding the nature of language. The second approach is classification-based sentiment analysis, also known as supervised classification. It builds a sentiment classifier using a train set that contains labeled texts or sentences and tests new texts using the classifier. Statistical and machine learning techniques can be used in this approach. The Bayesian modeling approach has proven to be a capable method for multi-class sentiment classification and multi-dimensional sentiment distribution predictions [22]. Machine learning techniques, such as Naïve Bayes (NB), Support Vector Machines (SVM), Maximum Entropy, Decision Tree and K-Nearest Neighbor Classifier have been shown to be effective methods for sentiment analysis of messages [23][24].

Some of sentiment analysis approaches examine documents' author information or behavior. Guerra et al. [25] proposed a sentiment analysis algorithm using the bias of social media users toward a topic. They posit users tend to express their opinion multiple times and a user's bias tends to be more consistent over time as a basic property of human behavior. Thus, they measured bias of social media users toward a topic and analyzed sentiment by transferring users biases into textual features. Kucuktunc et al. [26] also proposed a method of analyzing sentiment based on characteristics of users, such as gender, age and education level. However, these methods cannot be broadly used because it requires relationship data among users and previous text documents that the users have posted, which are not always provided by social networks because of the privacy laws.

Speriosu et al. [27] applied the label propagation (LPROP) approach based on graph representation to analyze the sentiment of documents in Twitter. Their assumption is that each tweet written by a user is linked to other tweets written by the same user, and each author is influenced by the tweets written by users whom he or she follows. They represented such a relationship using a graph where the features of the document, such as words, emoticons, and authors, are inter-related to each other. Those features affect positivity or negativity of the documents in the graph. They tested the accuracy of the LPROP approach with messages in four different topics and compared it with the accuracies of other approaches. The results show that accuracy of the proposed LPROP approach is the highest among other sentiment analysis approaches as it reached 65.7% to 84.7%, depending on the topics. However, there is a room for improving the accuracy of the LPROP since its average accuracy is still 72.08%.

In spite of high demands for automatic sentiment analysis on text documents in online social network data, the development of the automatic sentiment analysis faces some challenges as the text documents in online social networks are unstructured, unlabeled, dynamic and noisy [18][25]. Because of the characteristics of the text documents in social networks, the accuracy of previous automatic sentiment analysis approaches remains around 80%, which should be further improved for more accurate analysis. In addition, some existing approaches require additional information, such as user’s tendencies or relationships, which are not always available on online social networks.

2.4. Topic Model

The topic model uncovers abstract topics within text documents [28][29][30]. With the increasing popularity of online social networks, using the topic model for short texts documents has become an important task in social network analysis. The term frequency-inverse document frequency (TF-IDF) discovers how important a word is in a text document from corpus using statistical methods [31][32]. The importance of a word in a document is measured by two factors; term frequency and inverse document frequency. Term frequency tf(t,d) is the number of times that a term t occurs in a document d. Inverse document frequency idf(t,D) is measure of how much information the word provides, meaning whether the term is common or rare across all documents. Idf(t,D) is computed as shown in (1).

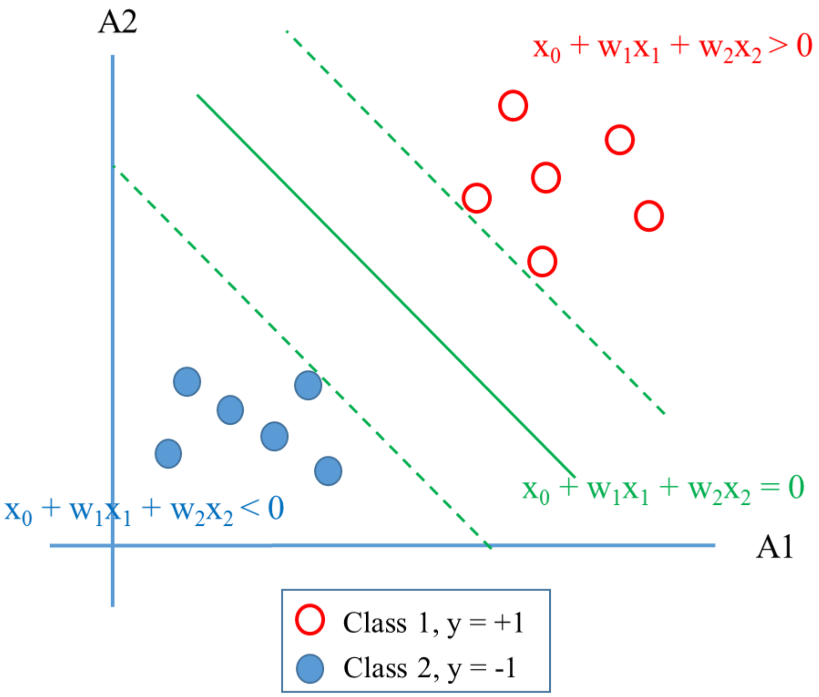
$$\mathrm{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \qquad (1)$$

In (1), d is a document in corpus D, and t is a term that appeared in a document d. N is the number of documents in the corpus D. $|\{d \in D : t \in d\}|$ is the number of documents where the terms t appears. Then, tdidf(t,d,D) is computed as shown in (2).

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \times \mathrm{idf}(t, D) \qquad (2)$$

The latent Dirichlet allocation (LDA) is a generative probabilistic framework that models documents as a finite mixture over an underlying set of topics [33]. In LDA, a document is considered a mixture of topics and represented as a distribution of words over the vocabulary. Those approaches are widely used in the topic model. However, LDA does not show high performance in short length text documents like it does in long length text documents.

In addition, classification algorithms can be used as a topic model to group similar documents. The Support vector machine (SVM) is a supervised learning method for classification and regression [34][35]. Given the train set where each row of data is classified as one two categories, SVM trains the features of each category and searches for a hyper lane that separates two groups in maximum margin. Then, it classifies new data as one of two categories using the maximum marginal hyper lane.


A2
$x_0 + w_1x_1 + w_2x_2 > 0$
$x_0 + w_1x_1 + w_2x_2 = 0$
$x_0 + w_1x_1 + w_2x_2 < 0$
A1
○ Class 1, y = +1
● Class 2, y = -1

- $D = \{(X_i, y_i)\}$
  - D : Classification dataset
  - X_i : the set of training tuples
  - y_i : class label (either +1, -1)
- $\boldsymbol{W}^*\boldsymbol{X}+b = 0$: Hyperlane
  - *W : a weight vector* $\{w_1, w_2,\ldots,w_n\}$
  - n : set of attributes
  - b : a scalar(bias)

Figure 1.Example of Classification using SVM

Figure 1 shows an example of classification using SVM. *D* is a classification data set where each tuples in the data set is classified as +1 or -1, and contains set of training features $X_n = \{x_1, x_2\}$ in it. Then, it generates the hyper lane, $x_0 + w_1x_1 + w_2x_2 = 0$, that separates the tuples labeled +1 from the other tuples labeled -1 with maximum margin. Given the new data tuples, it is classified as class +1 if the tuples lies above the separating hyperlane ($x_0 + w_1x_1 + w_2x_2 > 0$). In a similar way, the new data tuple is classified as class -1 if the tuples lies below the separating hyper lane.

The Naïve Bayes classifier is a probabilistic classifier based on Bayes' theorem and suitable for the data that contains high dimensionality [36][37][38]. In the Naïve Bayes classifier, *n* features for a data instance are represented in vector space model as shown in the (3) where *x* represents each feature in a feature set *X*.

$$X = (x_1, x_2, \dots, x_n) \tag{3}$$

Then, the probability of each data instance is assigned using conditional probability of each feature as shown in the (4). Then (4) is decomposed to (5) using Bayes' theorem

$$p(C_k|x_1, x_2, \dots, x_n) \tag{4}$$

$$p(C_k|\mathrm{X}) = \frac{p(C_k)\, p(\mathrm{X}|C_k)}{p(\mathrm{x})} \tag{5}$$

The Decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label [39][40]. It is one of the most popular classification methods based on feature-based learning [41]. Decision Tree induction is the learning of decision tree from class-labeled training tuples. Quinlan presented decision tree induction algorithms (Iterative Dichotomiser 3 and C4.5) [39][42]. Breiman et al. also presented a binary decision tree induction algorithm based on the statistical model [43]. A general decision tree induction algorithm is explained in Figure 2.

Input:

- Data partition D, which is a set of training tuples and their associated class labels;
- attribute_list, the et of candidate attributes;
- Attribute_selection_method, a procedure to determine the splitting criterion that that “best” partitions the data tuples into individual classes. This criterion consists of a splitting_attribute and, possibly, either a split-point or splitting subset.

Output: A decision tree

Method:

```
create a node N;
if tuples in D are all of the same class, C, then
    return N as a leaf node labeled with the class C;
if attribute_list is empty then
    return N as a leaf node labeld with the majority class in D;
apply Attribute_selection_method(D, attribute_list) to find the “best”
    splitting_criterion;
label node N with splitting_criterion
if splitting_attribute is discrete-valued and multiway splits allowed then
    attribute_list ← attribute_list – splitting_attribute;
for each outcome j of splitting_criterion
    let $D_j$ be the set of data tuples in D satisfying outcome j;
    if $D_j$ is empty then
        attach a leaf labeled with the majority class in D to node N;
    else attach the node returned by Generate_decision_tree($D_j$,
        attribute_list) to node N;
endfor
return N;
```

Figure 2. General decision tree induction algorithm [44]

2.5. Data Analysis Knowledge Extraction on Social Network Data

Extracting useful information and predicting a diverse variety of social phenomena is one of biggest issues in social network data analysis. Artis et al. proposed a stock market and box office forecasting model using Twitter data [45]. Relevance filtering, data cleaning, and sentiment analysis techniques were used for text pre-processing. They used the results of the text pre-processing as prediction features and several machine learning algorithms, such as Linear Regression, Neural Network, and SVM, were applied on them to predict the stock market index and box office data. Song et al. developed a Twitter data collecting and analyzing system to examine political data about the 2012 Korean presidential election on Twitter [1]. They collected 1,737,696 tweets that contain the name of presidential candidates and the keyword “Presidential Election” using the Twitter stream API. The tweets were pre-processed using such text analysis techniques. They analyzed the collected data using topic modeling, network analysis, and term co-occurrence analysis techniques. Mishne and Glance [46] extracted text documents written by bloggers and analyzed sentiment in them. The relationship between sentiment score and sales of 49 movies were examined by Pearson’s r-correlation. Asur and Hurberman [47] also analyzed sentiment on text documents and forecasted box-office revenues. They applied linear regression to predict sales of movies. Bollen et al. [48] estimated the general mood of tweets using number of tweets that contain positive or negative words, and mood status in profile. The estimated general mood were used as indicators for predicting Dow Jones Industrial Average (DJIA). Wolfram [49] directly extracted features from the text documents and predicted the NASDAQ stock prices. Predicting TV ratings and influenza rates using Twitter data has been also researched, and it showed promising results [50][51].

Table 3. List of Data Analysis and Prediction Model on Web Data

References	Data Analysis Target	Analysis / Prediction Model	Data
Arias et al.[45]	Stock Market Index, Movie Sales	SVM, Linear Regression, Neural Network	Tweets, Stock Market Index
Song et al.[1]	Korea Presidential Election	Topic Model, Network Analysis	Tweets
Mishne and Glance[46]	Movie Sales	Sentiment Analysis	Blog posts, IMDB Sales Data
Asur and Huberman [47]	Movie Sales	Linear Regression	Tweets
Bollen et al.[48]	Stock Market Index (DJIA)	SOFNN	Tweets, Stock Market Index
Wolfram[49]	Stock Market Index (NASDAQ)	SVM	Edinburgh Corpus, English, Relevant to stocks
Wakamiya et al.[50]	TV Ratings	Topic Model	Tweets
Lampos et al.[51]	Influenza Rates	Sparse Linear Regression	Tweets

3. Architecture and Process of Data Collection and Management System

In this section, we demonstrate architecture and implementation of the data collection and management system. The system consists of three main components, Crawler, Analyzer, and the Hadoop controller, where each component is inter-connected with each other. The overall system flow works as follows. First, the Crawler collects data from Twitter, YouTube, and the New York Times using their Application Programming Interface (API). Specifically, tweets from Twitter, video comments from YouTube, and news articles from New York Times are collected in this step. A tweet, comment and news article are considered documents in this system. Once the documents are collected by the Crawler, the Duplication checker filters duplicated documents and the Language detector identifies documents written in English. Second, the Analyzer pre-processes the documents using several text analysis techniques. Once the Term extractor retrieves meaningful terms using the Stanford core NLP [16], the Topic analyzer filters noisy data and extracts documents that related to a specific topic. Then, the Sentiment analyzer evaluates the polarity of each of the document, whether it contains either positive or negative opinion toward a topic. Finally, the Hadoop controller saves the number of documents, terms and sentiment score into HDFS (Hadoop distribute file system). We designed our own directory architecture for HDFS, which enabled us to process and manage text documents efficiently. The overall system architecture and document processing flows are depicted in Figure 3.

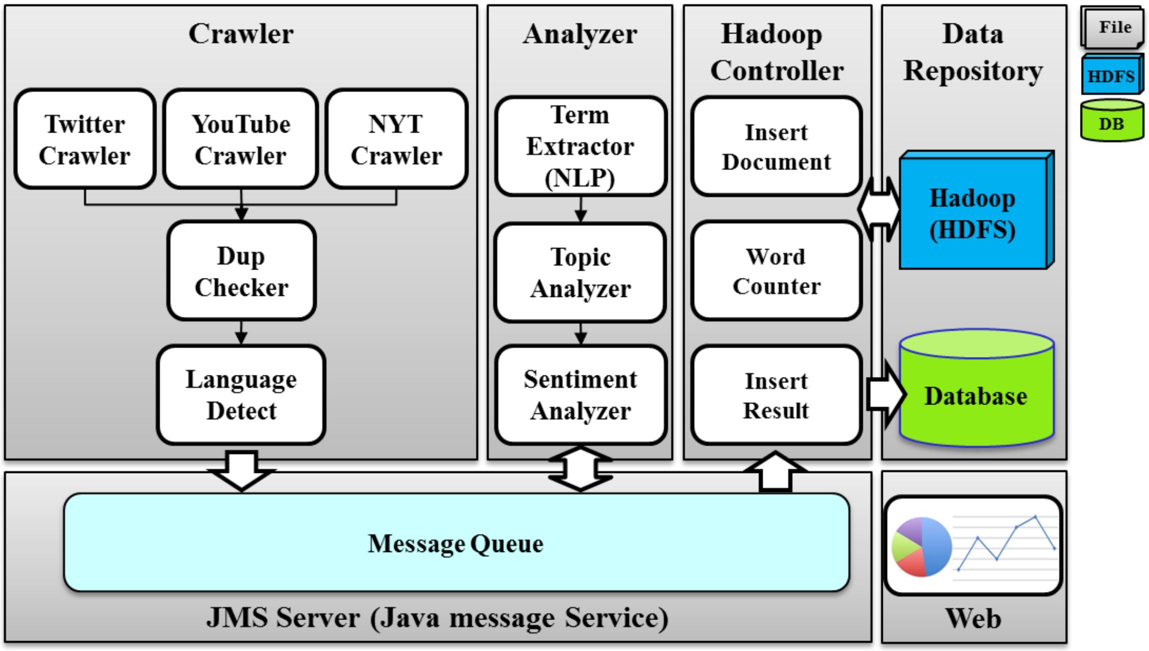


Figure 3. The System Architecture and Documents Processing Flows

In addition, we use a Java based message service server for communication channel that guarantees the independency of each component, further enable them to process in distributed computing. Details of each component and data processing are explained in the rest of this section.

3.1. Data Crawler

Previous research collected data from only single source channel, mostly Twitter as explained in Section 2.1. Nevertheless, users in a social network site might have bias toward a specific topic. In this research, we collected data from three source channels, Twitter, YouTube and New York Times for more accurate analysis of trends in real society.

Twitter crawler consists of seed and search crawler. The Seed crawler retrieves tweets from the selected Twitter accounts called seed nodes. The details of seed selection process are explained in Section 6. A Twitter account has unique screen name and numeric type Twitter account ID. This system only uses the Twitter account ID since the screen name can be changed by the own user although the account ID is permanently used in Twitter. All seed nodes information provided by seed selection process is stored in the database with the screen names and the Twitter account IDs. Once the seed node information is saved and ready to use, the Seed crawler sends a request to Twitter with the account id and receives the latest 3200 tweets by each seed node.

The Search crawler retrieves tweets using keywords that are related to a topic. The related keywords are manually inserted by human. After keywords are inserted into system, the Search crawler sends requests with the keywords using Twitter API. Then, Twitter will send response back to the system with the tweets containing the specific keywords. Normally, Twitter provides the tweets that generated in last nine days from the time the system requests.

YouTube also provides us with API that allows us to collect YouTube video related information, such as video title, the number of views, likes, dislikes, profile of users and comments posted on a video. The YouTube Crawler collects comments posted on specific videos using the API. The target videos can be manually selected by human or searched using the keywords. The New York Times crawler has been developed in the similar way as the YouTube Crawler. It sends requests to New York Times with keywords and receives the news articles related to keywords.

Once the Crawler collects the tweets in Twitter, comments in YouTube and the news article in New York times, each of them is treated a document in this system. All documents are sent to the duplication checker and it examines whether the same document already exists in the database or not. If the same document exists in the database, the new document is deleted in this step. Finally, the Language detector evaluates the each document and identifies the document that written in English. The documents not written in English are deleted in this step. We developed the Language detector based on Microsoft Language Detection Module. Figure 4 shows the architecture and process of the Data crawler. All processed documents are sent to the Message queue using the Message sender.

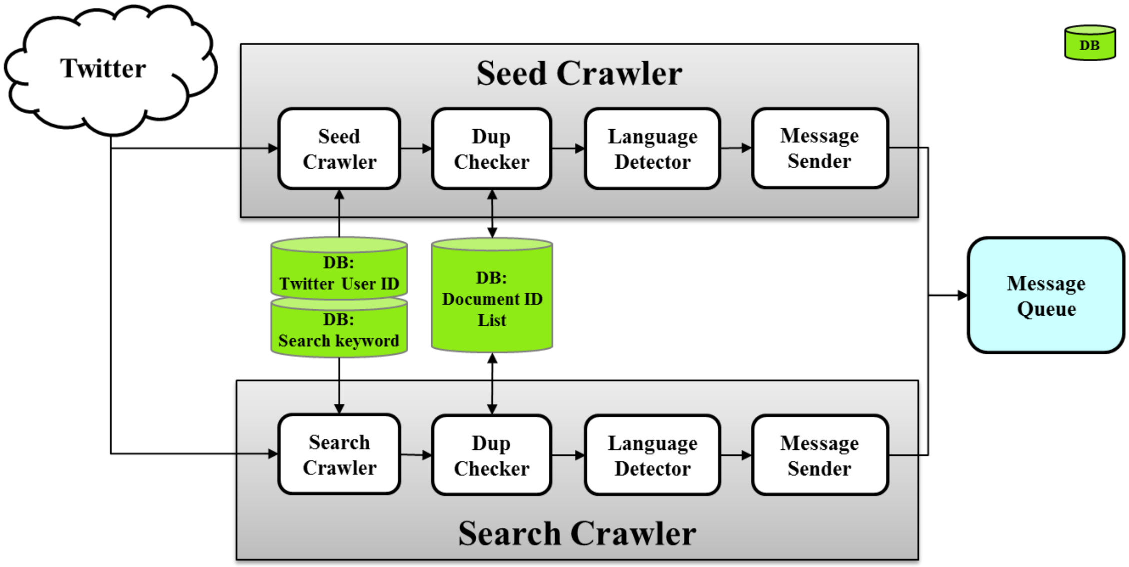


Figure 4. Architecture and Process of Twitter Crawler.

3.2. Analyzer

The Analyzer is a document pre-processor that extracts meaningful terms, filters out garbage data, identifies documents related to specific topics, and decides whether the documents are positive or negative opinion toward a topic. The architecture and general process of the Analyzer is depicted in Figure 5.

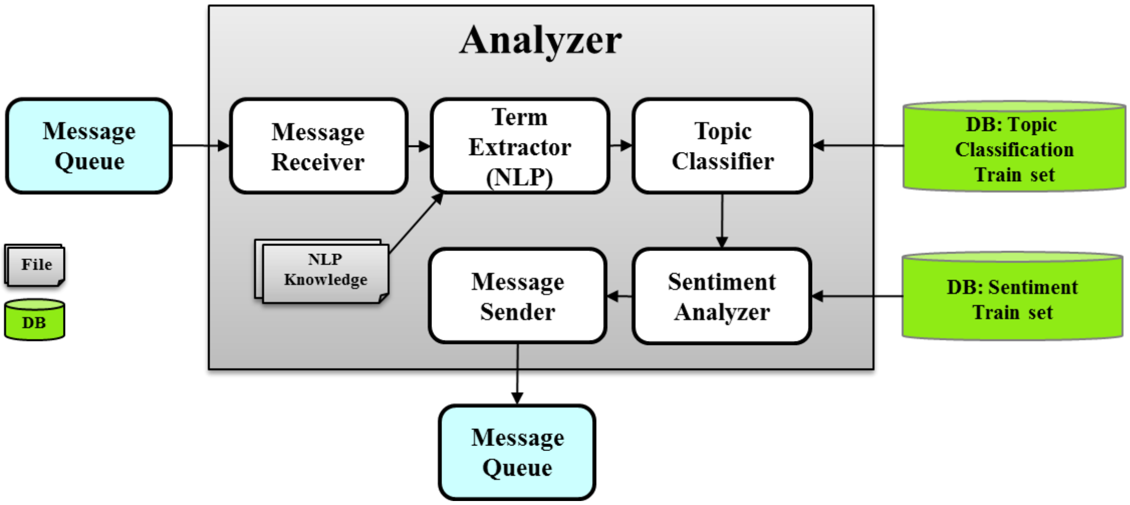


Figure 5. Architecture and Process of Analyzer

The Term Extractor discovers meaningful words in documents using NLP techniques. First, every word in a document is parsed into tokens and saved into a vector space. Second, the Term Extractor identifies morpheme of each word and labels it noun, adjective, verb or adverb, which is classified as a term. The words that are labeled other types of morpheme, such as pronouns and prepositions, are not considered terms since these types of morphemes do not have informational meaning. Finally, the original document and the extracted terms are saved into HDFS for further analysis. Additionally, it generates a list of the top 100 terms that most frequently appeared in documents by topic. The 100 terms are generated every day and saved into HDFS through the Map Reduce technique. We integrated Stanford Core NLP made by Stanford Natural Language Processing Group into the Term Extractor.

The Topic classifier retrieves documents that are related to specific topics and filters out the other documents that are not related to them. Terms for each document generated from the previous step are used as features to train the two documents set, documents related to the topic and documents not related to the topic. Then, it builds three text classifiers using SVM, Naïve Bayes and Decision Tree. Once a new document is given, the Voting module runs three classifiers and categorizes it into topic-related or not topic-related document. Categorization results are saved into HDFS with original documents using the Hadoop controller. Details of the classification process and experiment results are explained in Section 4.

The Sentiment analyzer extracts an opinion whether a document contains a positive or negative meaning toward a topic. We developed a sentiment analysis algorithm based on a probability model in previous research [52]. It reads sample documents in a train set and builds a sentiment lexicon that contains the list of words that appeared in the sample text documents and the probability that a text document is positive opinion if it includes these words. Then, it computes the positivity score of documents in a test set using the list of words in a document and sentiment lexicon. Each document is categorized as either positive or negative opinion, depending on the threshold value calculated using a train set. Details of sentiment analysis algorithm, accuracy and experiment results are explained in Section 5.

3.3. Hadoop Controller

The Hadoop controller stores original documents, terms and sentiment score into HDFS in a predefined HDFS file structure format. It also generates the most frequently used words by topic and saves the list of words every day. Figure 6 depicts the architecture and process of the Hadoop controller.

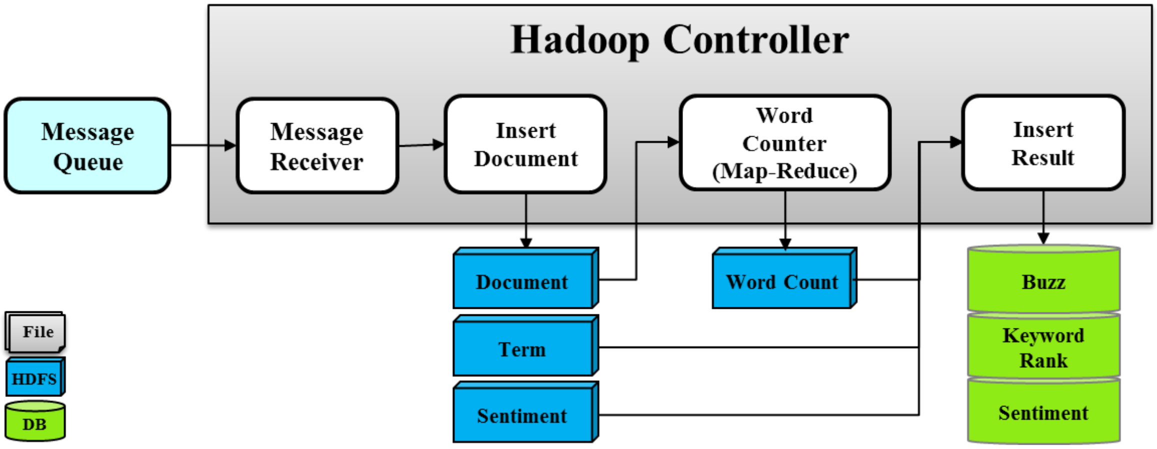


Figure 6. Architecture and Process of Hadoop Controller

We designed Hadoop-based file and directory architecture to save, analyze and retrieve documents efficiently as shown in the Figure 7. Every document is saved by channel, company, category and date. The documents, terms, and sentiment scores are stored in separate files in order to retrieve these data without additional processing.

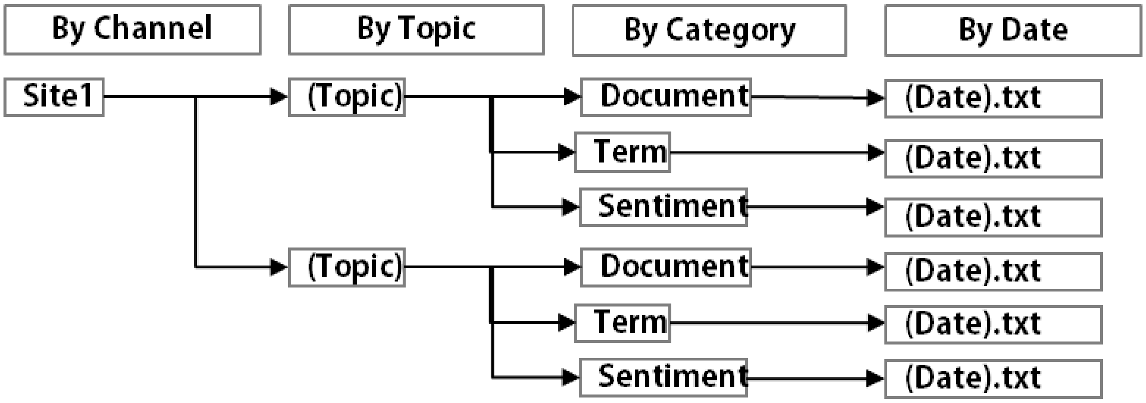


Figure 7. File and Directory Architecture in HDFS

The Word counter extracts the most frequently used words by topic and sorts them by the number of appearance. The list of words and number of appearances are saved into HDFS using Map-Reduce, which enable us to directly access to the data without additional processing.

3.4. Message Processor

In this system, all components are communicated with each other using a messaging service server that enables them to process data in distributed computing. The messaging service can be understood as an exchange of messages between software components such as TCP (Transmission Control Protocol) network socket, CORBA (Common Object Request Broker Architecture), or RMI (Java Remote Method Invocation). This approach allows software components to communicate indirectly using message queues. The main benefit of this approach is that a message sender does not necessarily need to know the status of a receiver in order to send a message. Also, it helps to integrate heterogeneous platforms, reduce system bottlenecks, increase scalability, and change functionality of a component not affecting the other components.

This system supports multi-processing as performance gaps exist among the crawler, the analyzer, and the Hadoop controller. Depending on the speed of each component, the system runs additional process to avoid bottle-neck problems. For examples, if the analyzer is about 7 times slower than the crawler, the system runs 7 times more processes for the analyzer than the crawler.

3.5. Data Collection Experiment Results

In these data collection experiments, we selected 22 companies listed in 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 Table 4.

Table 4. 22 Target Companies for Data Collection and Analysis

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 six Twitter app keys to the Seed crawler. The six app keys allow the Crawler to send 1,080 queries per 15 minutes. Since a query is able to collect 100 tweets, 108,000 tweets can be crawled every 15 minutes (7,200 tweets per minute). We also assigned 28 Twitter app keys to the Search crawler. The 28 app keys allow the Crawler to send 5,040 queries per 15 minutes, meaning 504,000 tweets can be crawled every 15 minutes (33,600 tweets per minute). Limitation of app keys for YouTube and New York Times are not described in their API documentation.

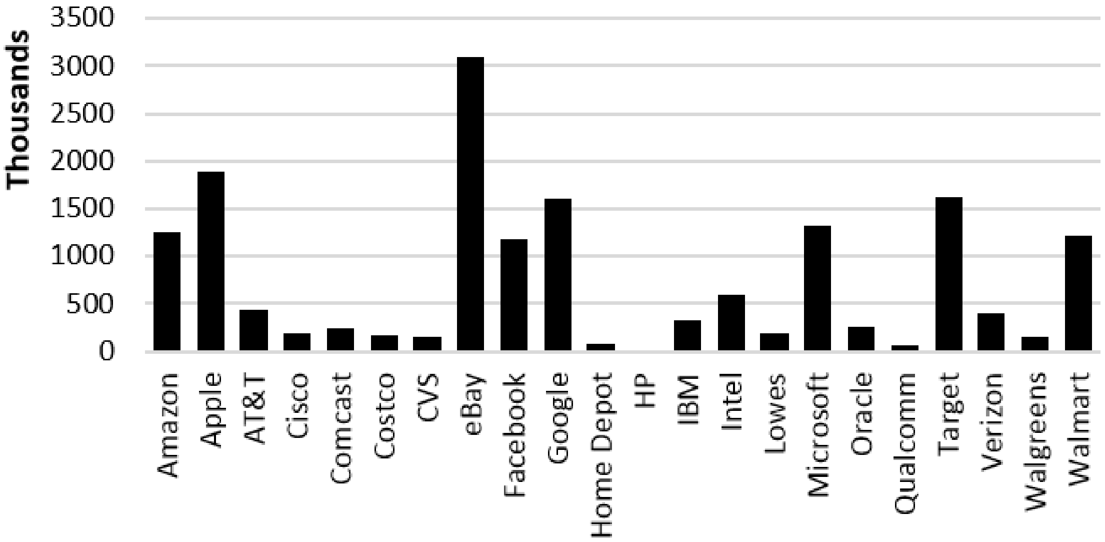


Figure 8. The Number of Documents by Companies

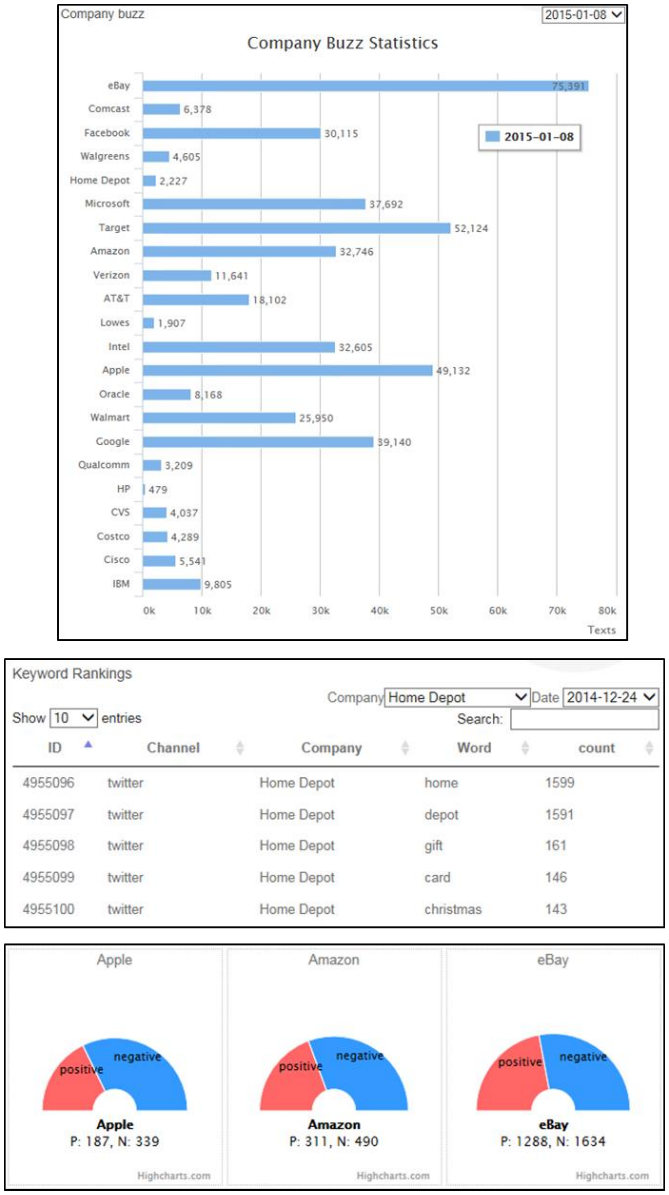
From Dec 19, 2014 to Jan 18, 2015, the crawler collected 16,479,483 documents. On average, it collects about 532,226 documents daily. We can expect that about 2 billion documents will be collected per a year using this system. Figure 8 shows the total amount of documents by the companies collected by this system.

Table 5 shows comparisons of performance among the Crawler, Analyzer and Hadoop controller. Numbers in a cell indicate the number of document processed in 10 minutes. On average, the Crawler processes documents 6.3 times faster than the Analyzer does as they handle approximately 3,177 and 504 tweets respectively. Also, the Crawler is 1.4 times faster than the Hadoop controller as shown in the Table 5. Therefore, the system allocates 7 times more processes to the Analyzer than Crawler. In the similar way, 2 times more processes are allocated to the Hadoop controller than the Crawler.

Table 5. Performance Comparisons among Crawler, Analyzer and Hadoop Controller

Processing 10 minutes	Crawler	Analyzer	Hadoop Controller
test 1	3,681	599	1,967
test 2	3,539	566	2,149
test 3	3,545	455	2,161
test 4	3,527	489	2,294
test 5	3,755	448	2,171
test 6	2,789	443	2,155
test 7	2,171	451	2,180
test 8	2,000	491	2,152
test 9	3,670	483	2,260
test 10	3,098	616	2,727
average	3,178	504	2,222

Moreover, we also developed a web application based on Java and Spring framework to report summaries of results by the companies and dates including the number of documents, word counts, and sentiments as shown in the Figure 9. The number of documents, frequently used words, and positive and negative documents are depicted in graph and table format for users to easily understand the results of summaries.



Keyword Rankings

Company Home Depot Date 2014-12-24

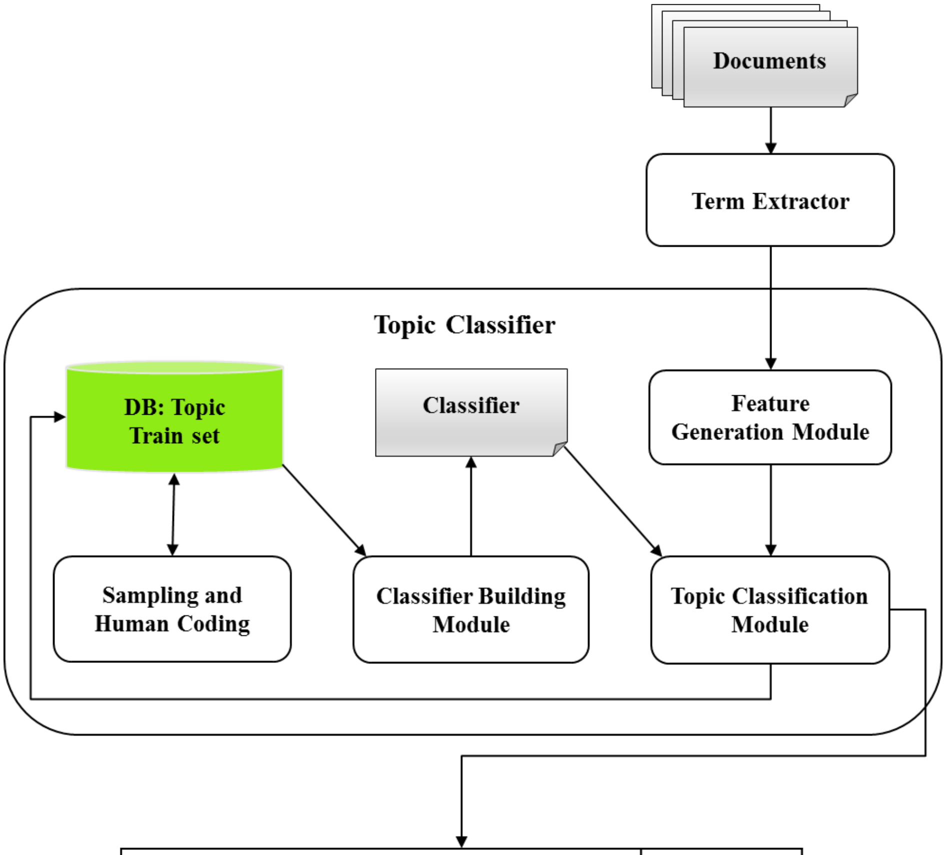
Show 10 entries

Search:

ID	Channel	Company	Word	count
4955096	twitter	Home Depot	home	1599
4955097	twitter	Home Depot	depot	1591
4955098	twitter	Home Depot	gift	161
4955099	twitter	Home Depot	card	146
4955100	twitter	Home Depot	christmas	143

Figure 9. Web Based User Interface for Summary of Data Processing Results

4. Topic Related Document Classification


Documents
Term Extractor
Topic Classifier
DB: Topic Train set
Classifier
Feature Generation Module
Sampling and Human Coding
Classifier Building Module
Topic Classification Module

Original Documents	Topic Classification Results(Ads)
HOT DEALS : http://t.co/4jouCa194e Apple iPhone 5 16GB a1428 (AT&T) White Black $199.99	Positive
I hate talking to at&t they want to do everything but send me a new phone.	Negative
Apple iPad mini 2 128GB Wi-Fi + 4G Cellular (AT&T) 7.9in - Space Gray - Full read by eBay: Price 335.0 USD (7? http://t.co/6lD4afgFKM	Positive
…	…

Figure 10. Process of Topic Classifier

In this section, we demonstrate the process and details of the implementations of the Topic classifier. The overall process of the Topic classifier is depicted in Figure 10. First, sample documents are retrieved from the raw data set labeled by human coders. For each topic, the documents are classified as a positive class if the documents are related to the topic. Otherwise, they are classified as a negative class. Second, the Classifier building module generates SVM, Naïve Bayes and decision tree classifiers for each topic. Once the new documents are given, the Term extracting module retrieves meaningful words from original documents as explained in Section 3.2, which is called *terms* in this research. Then, the Feature generation module creates a feature vector where existence of frequently used words and special topic related keywords are represented as Boolean or numeric values. Finally, The Topic classification module classifies documents based on the classifiers and features generated in the previous step. Additionally, the Topic classification module classifies new train set candidates and saves them into the data store. Details of methodologies and experiment results are explained in the rest of this section.

4.1. Data Preprocessing and Building Classifier

In this research, we retrieved 4,000 sample documents generated from online social networks or web sites. Every sample document is labeled a topic by human coders. Five graduate students were involved in the coding process. The documents are classified as advertising, opinion, stock related or miscellaneous documents as shown in the Table 6. Ambiguous documents that at least one coder classified as different topic are excluded from the sample data set. As a result, we have extracted total of 2,857 sample documents as shown in the Table 7.

Table 6. Example of Document Topic Classification by Human Coders

Documents	**Advertising**	**Opinion**	**Stock**	**Miscellaneous**
Apple iPhone 6 (Latest Model) - 64GB - Space Grey (AT&T) Smartphone via eBay http://t.co/LNgLvQTh3I	**Positive**	Negative	Negative	Negative
At&t got the worst customer service	Negative	**Positive**	Negative	Negative
@HuckNineteen Or Grandma's AT&T stock that she acquired in dribs and drabs. Exempts $250k per person for house, $100 k other.	Negative	Negative	**Positive**	Negative
#Dow #stocks $T AT&T Daily:0.14% Weekly:-0.4% YTD:3.33% Trend:64 / 100 http://t.co/gHyX10bVga	Negative	**Positive**	Negative	Negative
Great Deals #424 : http://t.co/7kT5iuJwIn Apple iPhone 5c - 16GB (AT&T) Smartphone - Blue - Pink - White - Yell... http://t.co/BxzHWz3O6J	**Positive**	Negative	Negative	Negative
I'm really hoping AT&T gets my new iPhone to me before Christmas	Negative	Negative	**Positive**	Negative
At&t/cingular Samsung Sgh-a117 Flip Cell Mobile Wireless Go Phone: At&t/cingular Samsung Sgh-a117 Flip Cel... http://t.co/nX6x7L62Ps	**Positive**	Negative	Negative	Negative
@iMSUCHATHREAT_ you at T shunica house	Negative	Negative	Negative	**Positive**
RT @TripleDTrader: Some interesting sell imbalances this morning: $PFE -278K, $KO -162K, $BA -79K, $T -122K, $VZ -179K, $DOW -158K, $BAC -1?	Negative	**Positive**	Negative	Negative
?@lizjs_: $T got me the cutest secret Santa present??	Negative	Negative	Negative	**Positive**

Table 7. Experiment Documents Set

Category	**Advertising**	**Opinion**	**Stock**	**Miscellaneous**	**Total**
Documents	1,096	616	408	737	2,857

For each topic category, sample documents are split into positive and negative class. For example, there were 1,096 documents classified into adverting topic in the previous step. For adverting topic, these 1,096 documents were labeled positive class and the other 1,761 documents that are classified into opinion, stock and miscellaneous topic were labeled negative class. Then, for each positive and negative class labeled documents, we extracted 100 documents that will be used for test set in this research. In the same way, opinion and stock related documents were also labeled positive and negative class, which were spilt again into train and test set. Table 8 shows the number of sample train and test set for each topic category.

Table 8. Number of Train and Test Set for Each Category

-	Category	Advertising	Opinion	Stock
Train Set	**Positive**	996	516	308
	Negative	1661	2141	2349
Test Set	**Positive**	100	100	100
	Negative	100	100	100

Once the train set is built, the Classifier building module generates the Topic classifiers using SVM, Naïve Bayes and Decision Tree as shown in the Figure 11. First, the Term extraction module reads documents in topic train set and extracts list of terms for each document as explained in Section 3.2. Second, the Feature generation module generates the list of 100 frequently used terms from all terms in the train set, which is used for key features to train topic related documents. For each document, the existence of the each term is represented as vector space model with a zero or one value. If a term exists in a document, value of the term element in the vector is set for one. Otherwise, a zero is assigned to it. Also, the existence of stock symbol, URL and a string that indicates currency format are extracted as additional features. An example of term and feature extraction from documents is depicted in Figure 12. Finally, it trains the positive/negative class documents using the features and three types of classification algorithm, SVM, Naïve Bayes and Decision tree. Then, it generates three classifiers and saves them as a file format for the classification process of test documents.

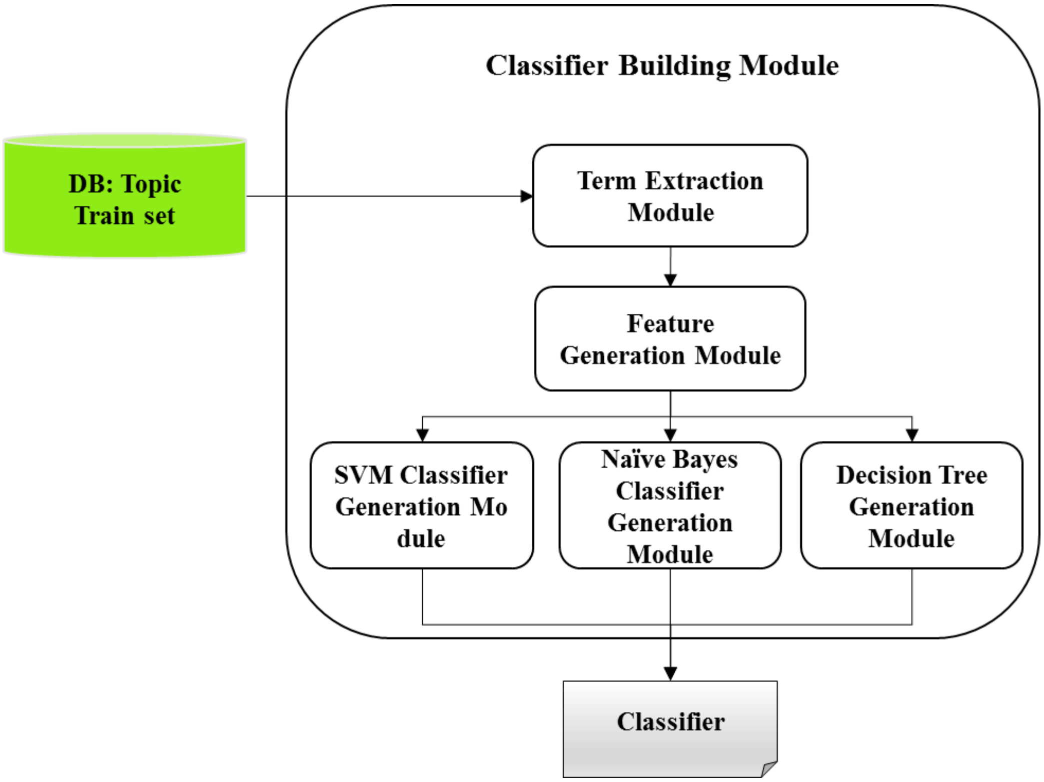
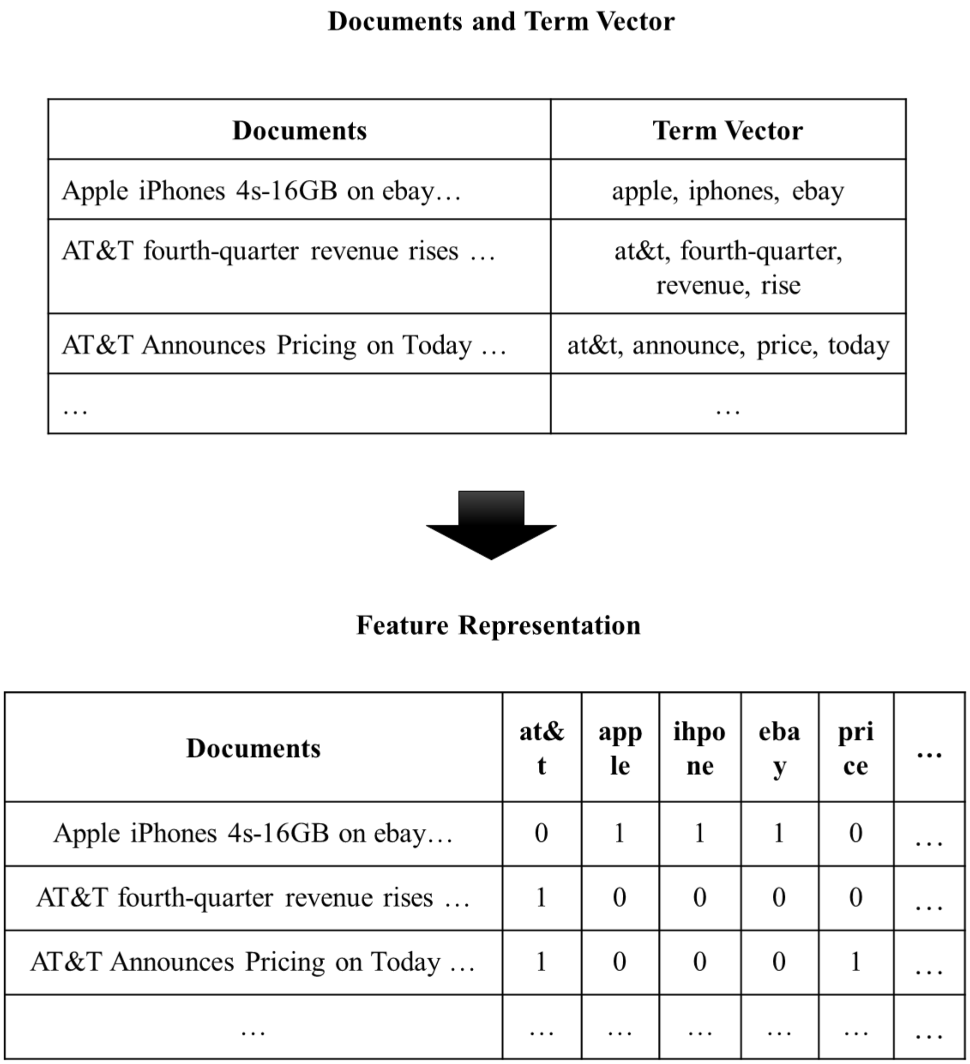

Classifier Building Module
DB: Topic Train set
Term Extraction Module
Feature Generation Module
SVM Classifier Generation Mo dule
Naïve Bayes Classifier Generation Module
Decision Tree Generation Module
Classifier

Figure 11. Process of Classifier Building Module



Documents and Term Vector

Documents	Term Vector
Apple iPhones 4s-16GB on ebay…	apple, iphones, ebay
AT&T fourth-quarter revenue rises …	at&t, fourth-quarter, revenue, rise
AT&T Announces Pricing on Today …	at&t, announce, price, today
…	…

Feature Representation

Documents	at&t	apple	ihpone	ebay	price	…
Apple iPhones 4s-16GB on ebay…	0	1	1	1	0	…
AT&T fourth-quarter revenue rises …	1	0	0	0	0	…
AT&T Announces Pricing on Today …	1	0	0	0	1	…
…	…	…	…	…	…	…

Figure 12. Term Extraction and Feature Representation Example

The SVM classifier is built as follows. Let document set *D* be a classification train data set with *n* documents as shown in the (6) , with *i* = 1, 2, …, *n*. Features of a document are represented in 103 dimensional space where 100 frequently used words and keywords “$number”, “$symbol”, and URL format indicate each feature. Then, let there be only two class labels such that y_i is either +1(positive class) or -1(negative class). A hyperlane *h(x)* in (7) draws a linear line that splits the original space into two half-space in 103 dimensional spaces. In this research, an R package, *e-1071*, has been used to build SVM classifiers.

$$D = \{(x_i, y_i)\} \tag{6}$$

$$h(x) = w_1x_1 + w_2x_2 + \cdots + w_{103}x_{103} + b \tag{7}$$

To build Naïve Bayes classifiers, we also used the e-1071 package and the same independent predictor variables (100 frequently used words and keywords "$number", "$symbol", URL format). Then it computes the conditional a-posterior probabilities of a categorical class variable given the independent predictor variables using the Bayes rule. Table 9 shows list of conditional probabilities for each independent predictor variables. Finally, the conditional probability for each document is computed as explained in (4) and (5) in Section 2.4.

Table 9. Probabilities of each features assigned by Naïve Bayes for Advertising Documents

	at&t	apple	iphone	ebay	Smartphone	Price	16gb	…	URL
p(positive)	0.94	0.48	0.44	0.43	0.39	0.37	0.36	…	0.99
p(negative)	0.52	0.01	0.01	0	0.01	0.04	0	…	0.6

The decision tree is induced by an R package named *rpart* and uses the same features as SVM and Naïve Bayes classifiers. We built three decision trees for each topic. Figure 13 shows an example of decision tree to classify advertising documents.

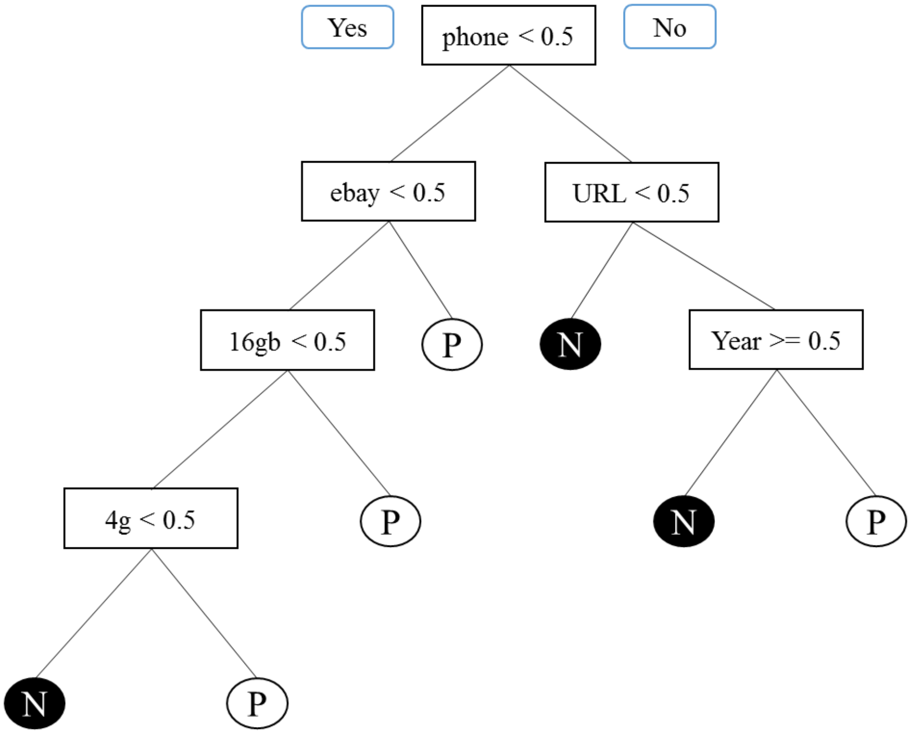
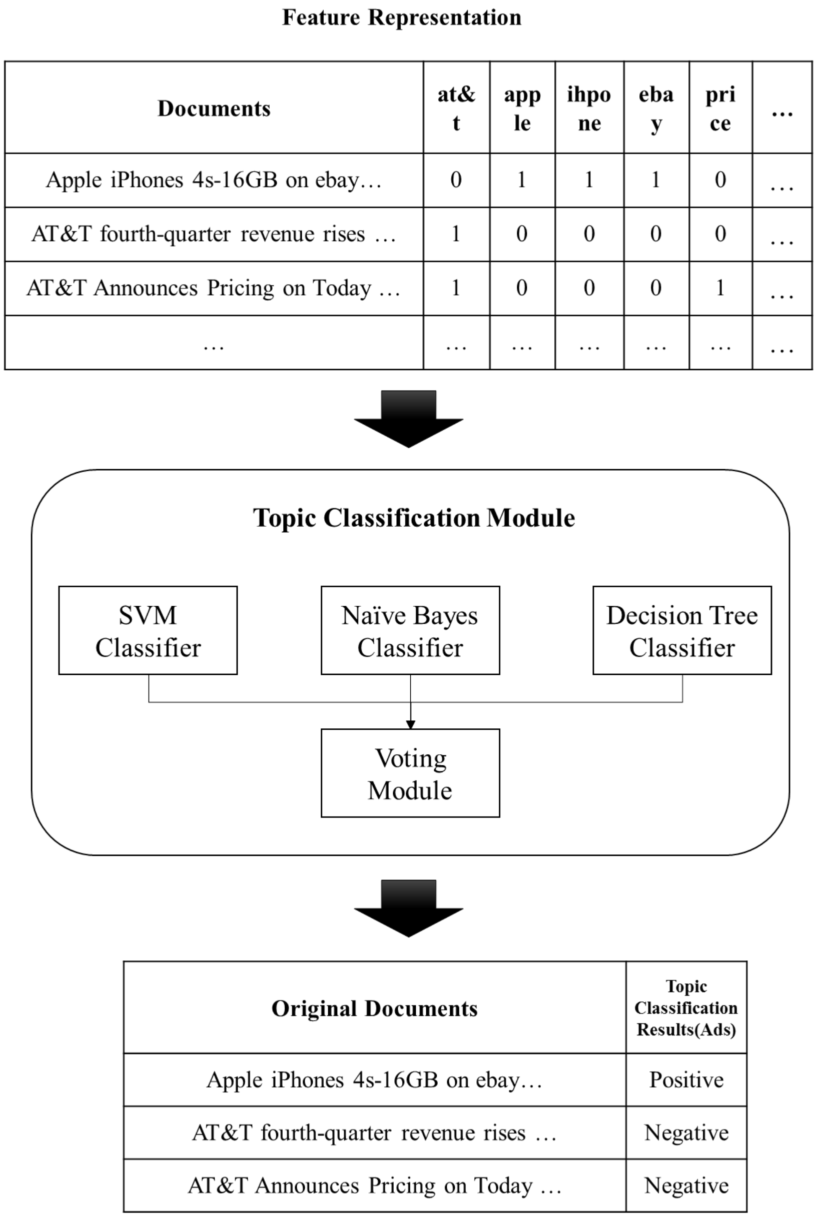


Figure 13. Example of Decision Tree to Classify Adverting Documents

4.2. Document Classification using Support Vector Machine, Naïve Bayes and Decision Tree

Once the three classifiers are built for each topic, the Topic classification module categorizes new documents using the classifiers. Then, Voting module finally classifies them based on the results derived by the classifiers. Figure 14 shows the overall process of the Topic classification module.



Feature Representation

Documents	at&t	apple	ihpone	ebay	price	…
Apple iPhones 4s-16GB on ebay…	0	1	1	1	0	…
AT&T fourth-quarter revenue rises …	1	0	0	0	0	…
AT&T Announces Pricing on Today …	1	0	0	0	1	…
…	…	…	…	…	…	…

Original Documents	Topic Classification Results(Ads)
Apple iPhones 4s-16GB on ebay…	Positive
AT&T fourth-quarter revenue rises …	Negative
AT&T Announces Pricing on Today …	Negative

Figure 14. Overall Process of Topic Classification Module

The hyper lane function in (7) is applied to the new documents with the same feature format as train set. It generates decision values for the documents that used for classification threshold. If it is positive value, the document is classified into a positive class. Otherwise, it is classified into a negative class. Table 10 shows example of document topic classification using SVM decision value. The Naive Bayes classifier computes the conditional probability for each document as explained in (4) and (5) in Section 2.4 for each document. If *p*(positive) is greater than *p*(negative), the document is classified into the positive class of the topic. Otherwise, it is classified into the negative class of the topic. Table 11 shows Naïve Bayes probabilities and classification example. Decision tree classification is processed by features generated in the previous step. Table 12 shows example of the features and classification results by the decision tree.

Table 10. SVM Decision Values and Classification Example

Original Label (Ads)	**Documents**	**Decision Values**	**SVM Classificati on Results**
Positive	Apple iPhone 4s - 16GB - Black (AT&T) Smartphone (MC918LL/A) (unlocked) - Full read by eBay: Price 68.0 USD (0? http://t.co/vgXkxhAont	2.47395455	Positive
Positive	Daily Deals: Unlocked 3.5" Android 4.4 Smartphone 3G WiFi GPS AT&T Straight Talk Cell Phone: $48.9... http://t.co/WMbfpUs4iQ #ebaydeals	1.10396221	Positive
Positive	NEW on EBAY: Apple iPhone 6 Plus (6+) 64GB AT&T Phone In White/Gold http://t.co/k0bfqm41V5 http://t.co/kSqQCalUPz	1.23073380	Positive
Positive	http://t.co/V8MXNLwnt9 #Deals #6128 Samsung Galaxy S 4 S4 Zoom C105a AT&T Unlocked 4G Android SmartPhone White ... http://t.co/hsKKUBjgdE	1.29783652	Positive
Positive	RT @ATTCares: @paachhecoo Stay relaxed Guadalupe! Order the phone you crave on AT&T Next at http://t.co/31rTyfvJn4 today! ^Kim https://t.c?	1.12522636	Positive
Negative	This screen protector AT&T sold me for my 6 plus is pure trash. Screen protector already cracking. I want answers @ATT	-1.57803923	Negative
Negative	@ognayah_ i hate at&t ??	-2.72114057	Negative
Negative	#sharknado is very very interesting, gotta love flying sharks	-2.72114057	Negative
Negative	$DVY Percent Change Updated Thursday February 12	-1.12082525	Negative
Negative	AT&T's Donovan: When it comes to 5G, timing is everything http://t.co/kghJH9eU6m	-1.00008603	Negative

Table 11. Naïve Bayes Probabilities and Classification Example

Original Label (Ads)	Documents	*p*(negative)	*p*(positive)	SVM Classification Results
Positive	Apple iPhone 4s - 16GB - Black (AT&T) Smartphone (MC918LL/A) (unlocked) - Full read by eBay: Price 68.0 USD (0? http://t.co/vgXkxhAont	0	1	Positive
Positive	Daily Deals: Unlocked 3.5" Android 4.4 Smartphone 3G WiFi GPS AT&T Straight Talk Cell Phone: $48.9... http://t.co/WMbfpUs4iQ #ebaydeals	0	1	Positive
Positive	NEW on EBAY: Apple iPhone 6 Plus (6+) 64GB AT&T Phone In White/Gold http://t.co/k0bfqm41V5 http://t.co/kSqQCalUPz	0	1	Positive
Positive	http://t.co/V8MXNLwnt9 #Deals #6128 Samsung Galaxy S 4 S4 Zoom C105a AT&T Unlocked 4G Android SmartPhone White ... http://t.co/hsKKUBjgdE	0	1	Positive
Positive	RT @ATTCares: @paachhecoo Stay relaxed Guadalupe! Order the phone you crave on AT&T Next at http://t.co/31rTyfvJn4 today! ^Kim https://t.c?	0.99	2.49e-04	Negative
Negative	This screen protector AT&T sold me for my 6 plus is pure trash. Screen protector already cracking. I want answers @ATT	1	1.89e-43	Negative
Negative	@ognayah_ i hate at&t ??	1	3.92e-40	Negative
Negative	#sharknado is very very interesting, gotta love flying sharks	1	3.92e-40	Negative
Negative	$DVY Percent Change Updated Thursday February 12	1	2.81-e131	Negative
Negative	AT&T's Donovan: When it comes to 5G, timing is everything http://t.co/kghJH9eU6m	1	6.82e-23	Negative

Table 12. Decision Tree Features and Classification Example

Original Label (Ads)	Documents	phone	ebay	…	URL	Decision Tree Classification Results
Positive	Apple iPhone 4s - 16GB - Black (AT&T) Smartphone (MC918LL/A) (unlocked) - Full read by eBay: Price 68.0 USD (0? http://t.co/vgXkxhAont	0	1	…	1	Positive
Positive	Daily Deals: Unlocked 3.5" Android 4.4 Smartphone 3G WiFi GPS AT&T Straight Talk Cell Phone: $48.9... http://t.co/WMbfpUs4iQ #ebaydeals	1	1	…	1	Positive
Positive	NEW on EBAY: Apple iPhone 6 Plus (6+) 64GB AT&T Phone In White/Gold http://t.co/k0bfqm41V5 http://t.co/kSqQCalUPz	0	1	…	1	Positive
Positive	http://t.co/V8MXNLwnt9 #Deals #6128 Samsung Galaxy S 4 S4 Zoom C105a AT&T Unlocked 4G Android SmartPhone White ... http://t.co/hsKKUBjgdE	1	0	…	1	Positive
Positive	RT @ATTCares: @paachhecoo Stay relaxed Guadalupe! Order the phone you crave on AT&T Next at http://t.co/31rTyfvJn4 today! ^Kim https://t.c?	1	0	…	1	Positive
Negative	This screen protector AT&T sold me for my 6 plus is pure trash. Screen protector already cracking. I want answers @ATT	0	0	…	0	Negative
Negative	@ognayah_ i hate at&t ??	0	0	…	0	Negative
Negative	#sharknado is very very interesting, gotta love flying sharks	0	0	…	0	Negative
Negative	$DVY Percent Change Updated Thursday February 12	0	0	…	0	Negative
Negative	AT&T's Donovan: When it comes to 5G, timing is everything http://t.co/kghJH9eU6m	0	0	…	1	Negative

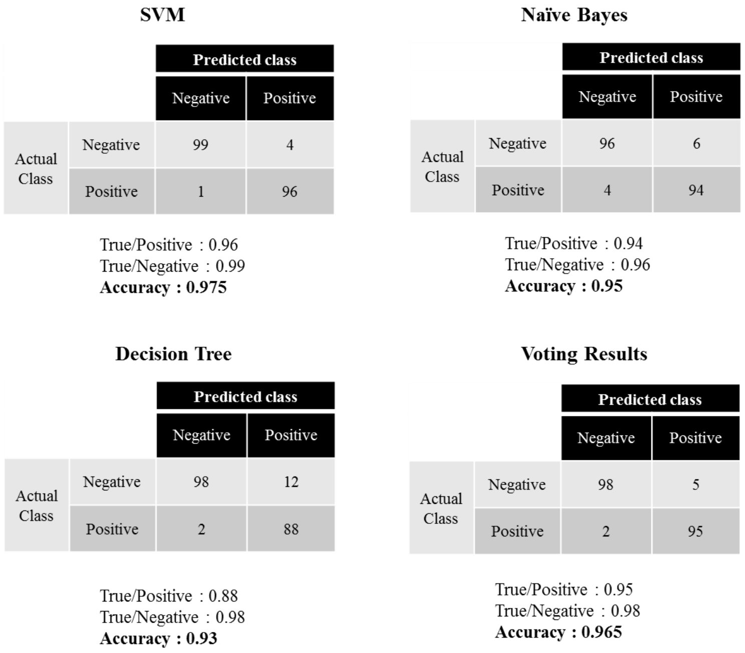
Table 13. Advertising Documents Voting Results and Classification Example

Original Label (Ads)	Documents	SVM	Naïve Bayes	Decision Tree	Voting Results
Positive	Apple iPhone 4s - 16GB - Black (AT&T) Smartphone (MC918LL/A) (unlocked) - Full read by eBay: Price 68.0 USD (0? http://t.co/vgXkxhAont	Positive	Positive	Positive	Positive
Positive	Daily Deals: Unlocked 3.5" Android 4.4 Smartphone 3G WiFi GPS AT&T Straight Talk Cell Phone: $48.9... http://t.co/WMbfpUs4iQ #ebaydeals	Positive	Positive	Positive	Positive
Positive	NEW on EBAY: Apple iPhone 6 Plus (6+) 64GB AT&T Phone In White/Gold http://t.co/k0bfqm41V5 http://t.co/kSqQCalUPz	Positive	Positive	Positive	Positive
Positive	http://t.co/V8MXNLwnt9 #Deals #6128 Samsung Galaxy S 4 S4 Zoom C105a AT&T Unlocked 4G Android SmartPhone White ... http://t.co/hsKKUBjgdE	Positive	Positive	Positive	Positive
Positive	RT @ATTCares: @paachhecoo Stay relaxed Guadalupe! Order the phone you crave on AT&T Next at http://t.co/31rTyfvJn4 today! ^Kim https://t.c?	Positive	Negative	Positive	Positive
Negative	This screen protector AT&T sold me for my 6 plus is pure trash. Screen protector already cracking. I want answers @ATT	Negative	Negative	Negative	Negative
Negative	@ognayah_ i hate at&t ??	Negative	Negative	Negative	Negative
Negative	#sharknado is very very interesting, gotta love flying sharks	Negative	Negative	Negative	Negative
Negative	$DVY Percent Change Updated Thursday February 12	Negative	Negative	Negative	Negative
Negative	AT&T's Donovan: When it comes to 5G, timing is everything http://t.co/kghJH9eU6m	Negative	Negative	Negative	Negative

Once classification results by SVM, Naïve Bayes and Decision tree are given, the Voting module derives final classification results by a majority voting system. For example, if more than two classifiers vote for positive class, the document is classified into the positive class. Otherwise, it is classified into the negative class. Table 13 shows the example of advertising documents classification. Furthermore, the system inserts documents into the train set if all three classification algorithm vote for the same class.

4.3. Topic Classification Experiment Results

We have conducted tree types of topic classification experiments: advertising, opinion and stock related documents as explained in section 4.1. Accuracies of SVM, Naïve Bayes, Decision tree and the final classification results by voting are derived for each topic as shown in the Figure 15, 16 and 17. Performances of each classification algorithm vary depending on the topic, as SVM shows the highest accuracy in advertising and opinion related documents, and decision tree shows the highest accuracy in stock related documents. The accuracy of the classification results by voting ranges 0.86 to 0.97. It is equal or greater than other classification algorithm in opinion and stock related documents. In adverting related documents, the accuracy of voting is 0.005 less than the highest accuracy among the three algorithms. Overall, performance of the proposed method is promising in that the accuracy shows 0.93 at average.



SVM

		Predicted class	
		Negative	Positive
Actual Class	Negative	99	4
	Positive	1	96

True/Positive : 0.96
True/Negative : 0.99
Accuracy : 0.975

Naïve Bayes

		Predicted class	
		Negative	Positive
Actual Class	Negative	96	6
	Positive	4	94

True/Positive : 0.94
True/Negative : 0.96
Accuracy : 0.95

Decision Tree

		Predicted class	
		Negative	Positive
Actual Class	Negative	98	12
	Positive	2	88

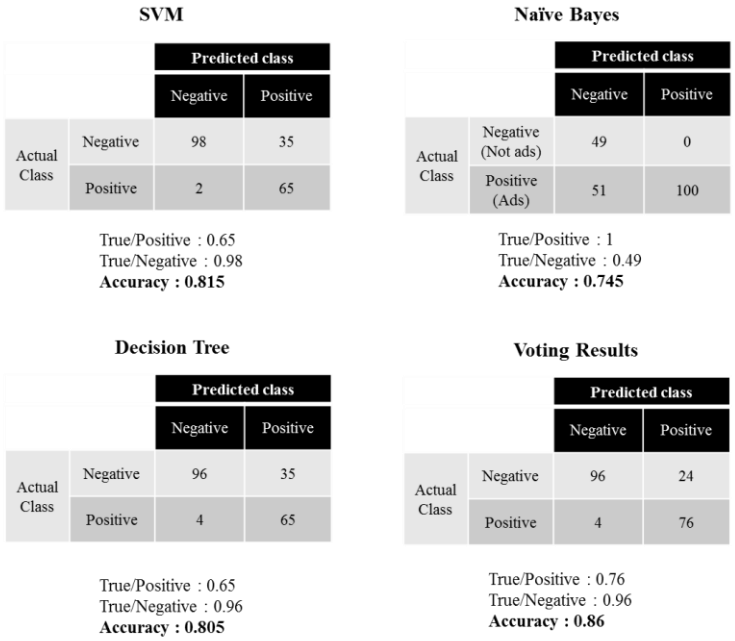
True/Positive : 0.88
True/Negative : 0.98
Accuracy : 0.93

Voting Results

		Predicted class	
		Negative	Positive
Actual Class	Negative	98	5
	Positive	2	95

True/Positive : 0.95
True/Negative : 0.98
Accuracy : 0.965

Figure 15. Accuracy of Advertising Related Documents



SVM

		Predicted class	
		Negative	Positive
Actual Class	Negative	98	35
	Positive	2	65

True/Positive : 0.65
True/Negative : 0.98
Accuracy : 0.815

Naïve Bayes

		Predicted class	
		Negative	Positive
Actual Class	Negative (Not ads)	49	0
	Positive (Ads)	51	100

True/Positive : 1
True/Negative : 0.49
Accuracy : 0.745

Decision Tree

		Predicted class	
		Negative	Positive
Actual Class	Negative	96	35
	Positive	4	65

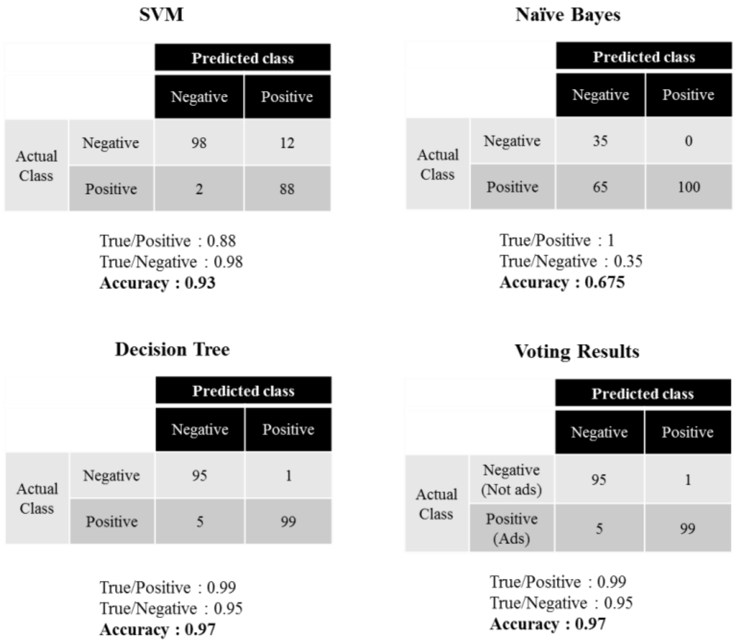
True/Positive : 0.65
True/Negative : 0.96
Accuracy : 0.805

Voting Results

		Predicted class	
		Negative	Positive
Actual Class	Negative	96	24
	Positive	4	76

True/Positive : 0.76
True/Negative : 0.96
Accuracy : 0.86

Figure 16. Accuracy of Opinion Related Documents



SVM

		Predicted class	
		Negative	Positive
Actual Class	Negative	98	12
	Positive	2	88

True/Positive : 0.88
True/Negative : 0.98
Accuracy : 0.93

Naïve Bayes

		Predicted class	
		Negative	Positive
Actual Class	Negative	35	0
	Positive	65	100

True/Positive : 1
True/Negative : 0.35
Accuracy : 0.675

Decision Tree

		Predicted class	
		Negative	Positive
Actual Class	Negative	95	1
	Positive	5	99

True/Positive : 0.99
True/Negative : 0.95
Accuracy : 0.97

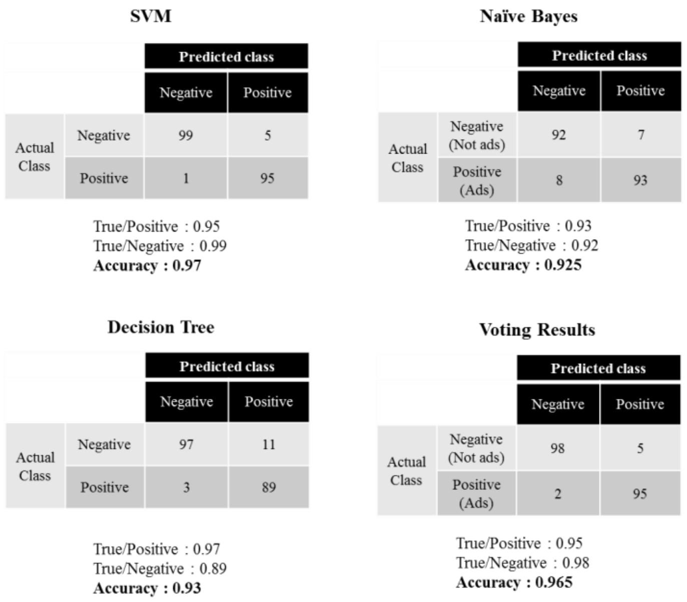
Voting Results

		Predicted class	
		Negative	Positive
Actual Class	Negative (Not ads)	95	1
	Positive (Ads)	5	99

True/Positive : 0.99
True/Negative : 0.95
Accuracy : 0.97

Figure 17. Accuracy of Stock Related Documents

We checked the accuracies of the algorithms again after the train set was automatically added. The same numbers of documents as in the initial train set are added into the train set and the classifiers are re-built by the new train set. Also, we used the same test set as the first experiment to compare the accuracies. The accuracies of classification algorithms are slightly different from the first experiment as shown in Figure 18, 19 and 20. However, the accuracies of classification results by voting are the same as the first experiment in all topics. These results indicate that documents that are automatically added into the train set are similar to existing documents in the train set. For this reason, even if the train set is added by algorithm, the accuracy of algorithm with new train set remains almost same as the accuracy with the initial train set.



SVM

		Predicted class	
		Negative	Positive
Actual Class	Negative	99	5
	Positive	1	95

True/Positive : 0.95
True/Negative : 0.99
Accuracy : 0.97

Naïve Bayes

		Predicted class	
		Negative	Positive
Actual Class	Negative (Not ads)	92	7
	Positive (Ads)	8	93

True/Positive : 0.93
True/Negative : 0.92
Accuracy : 0.925

Decision Tree

		Predicted class	
		Negative	Positive
Actual Class	Negative	97	11
	Positive	3	89

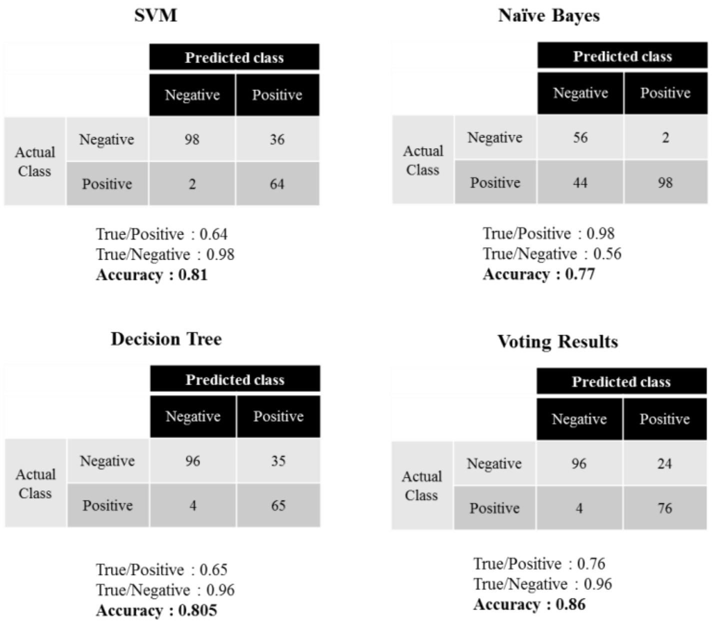
True/Positive : 0.97
True/Negative : 0.89
Accuracy : 0.93

Voting Results

		Predicted class	
		Negative	Positive
Actual Class	Negative (Not ads)	98	5
	Positive (Ads)	2	95

True/Positive : 0.95
True/Negative : 0.98
Accuracy : 0.965

Figure 18. Accuracy of Advertising Related Documents with Automatically Added Train Set



SVM

		Predicted class	
		Negative	Positive
Actual Class	Negative	98	36
	Positive	2	64

True/Positive : 0.64
True/Negative : 0.98
Accuracy : 0.81

Naïve Bayes

		Predicted class	
		Negative	Positive
Actual Class	Negative	56	2
	Positive	44	98

True/Positive : 0.98
True/Negative : 0.56
Accuracy : 0.77

Decision Tree

		Predicted class	
		Negative	Positive
Actual Class	Negative	96	35
	Positive	4	65

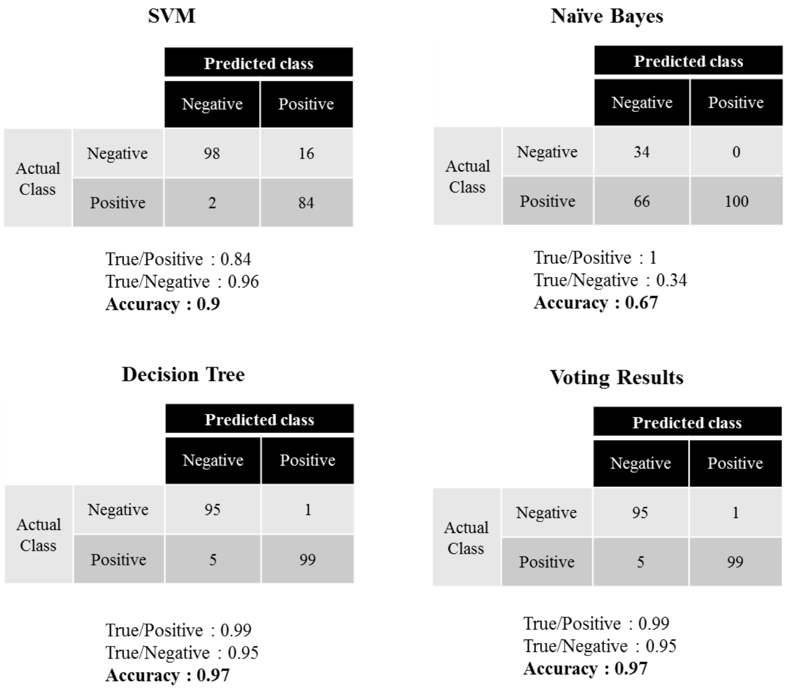
True/Positive : 0.65
True/Negative : 0.96
Accuracy : 0.805

Voting Results

		Predicted class	
		Negative	Positive
Actual Class	Negative	96	24
	Positive	4	76

True/Positive : 0.76
True/Negative : 0.96
Accuracy : 0.86

Figure 19. Accuracy of Opinion Related Documents with Automatically Added Train Set



SVM

		Predicted class	
		Negative	Positive
Actual Class	Negative	98	16
	Positive	2	84

True/Positive : 0.84
True/Negative : 0.96
Accuracy : 0.9

Naïve Bayes

		Predicted class	
		Negative	Positive
Actual Class	Negative	34	0
	Positive	66	100

True/Positive : 1
True/Negative : 0.34
Accuracy : 0.67

Decision Tree

		Predicted class	
		Negative	Positive
Actual Class	Negative	95	1
	Positive	5	99

True/Positive : 0.99
True/Negative : 0.95
Accuracy : 0.97

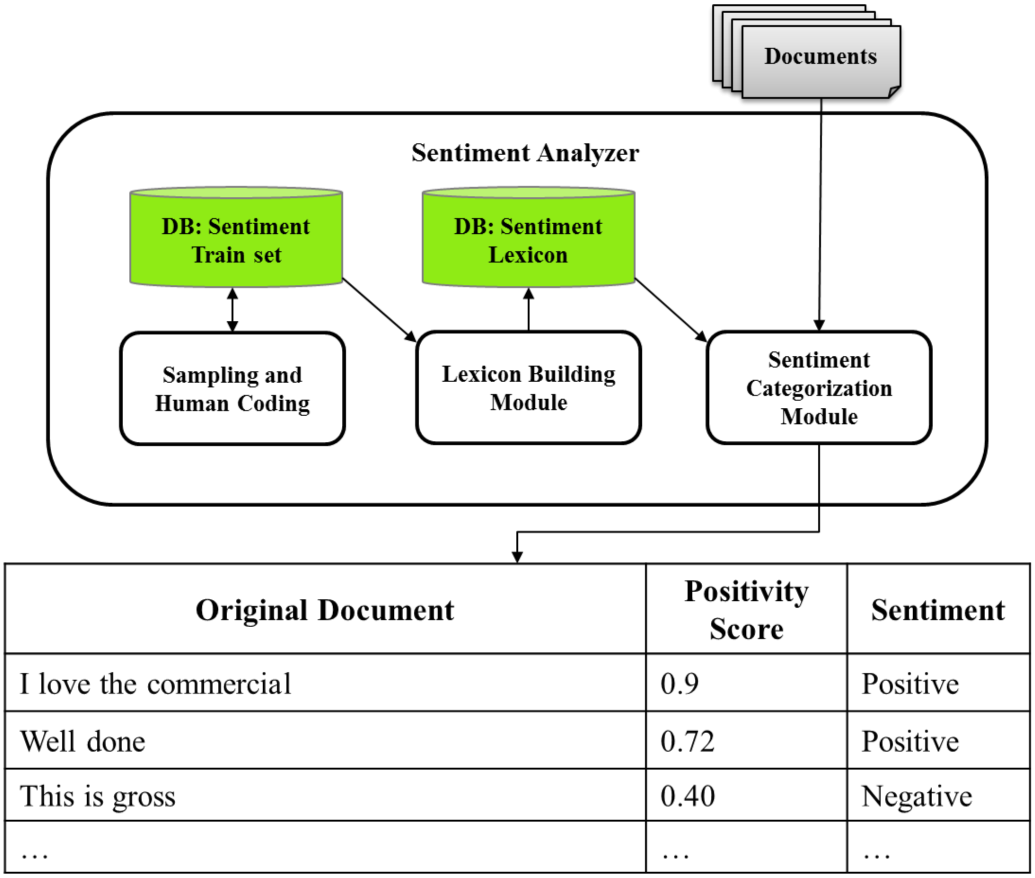
Voting Results

		Predicted class	
		Negative	Positive
Actual Class	Negative	95	1
	Positive	5	99

True/Positive : 0.99
True/Negative : 0.95
Accuracy : 0.97

Figure 20. Accuracy of Stock Related Documents with Automatically Added Train Set

5. Sentiment Analysis Using a Probability Model



Original Document	Positivity Score	Sentiment
I love the commercial	0.9	Positive
Well done	0.72	Positive
This is gross	0.40	Negative
…	…	…

Figure 21. Process of Sentiment Analyzer

In this section, we describe the methodology and implementation details of the Sentiment analyzer. Figure 21 shows the overall process of sentiment analysis on documents. First, *sample* documents for building train set are extracted from raw data set. The *sample* documents in the train set are categorized into positive or negative opinions by human coders. The categorized *sample* documents are saved into the train set with the sentiment label. After that, the Lexicon building module scans all categorized *sample* documents in the train set and calculates the weighted probability that the document is positive opinion if the word is included in a document. The list of words and the probabilities for each of them are saved in sentiment lexicon. Finally, the Sentiment categorization module calculates positivity scores for every document and categorizes whether the documents are positive or negative opinion. To check the accuracy of the proposed method, we generated a test set which is also categorized in the same way the train set is made. Details of methodologies are explained in rest of this section

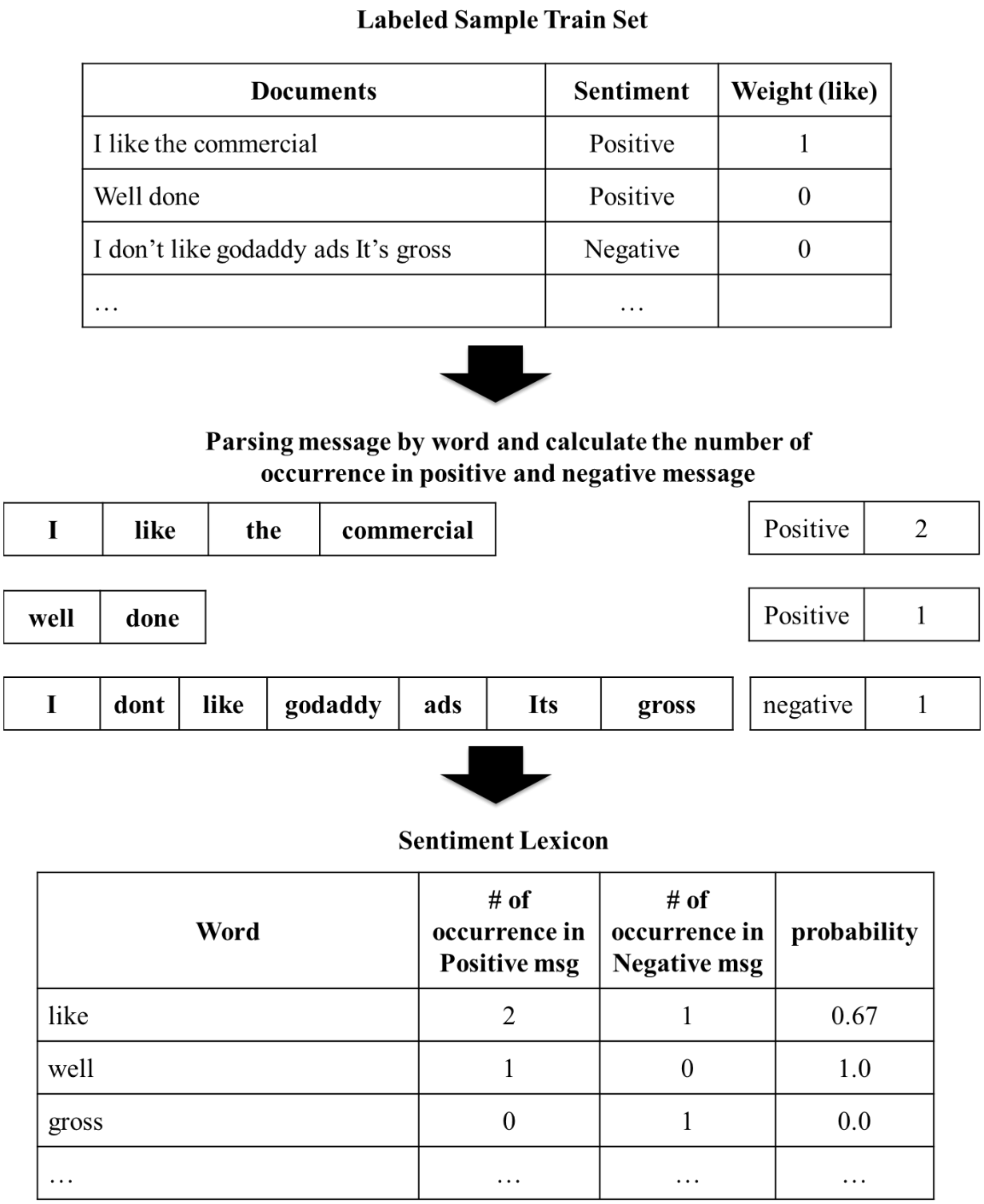
5.1. Sampling and Human Coding

For this research, we collected comments posted on three YouTube videos : *Prom* (for Audi), *Farmer* (for Ram) and *Perfect match* (for Go Daddy) that aired during the Super Bowl Game in 2013 which created a lot of buzz on online social networks. Among the all comments, we randomly selected a total of 3,000 comments, 1,000 comments for each video. A comment posted on a YouTube video is considered a document in this system. The documents were categorized as positive or negative opinions by human coders. Two graduate students were involved in the coding process. We built a data sample using the documents that both human coders categorized into the same sentiment. In this process, we excluded documents that have neutral or mixed opinions that have both positive and negative opinions in the sample documents. The categorized documents are saved into a train set in the data store.

5.2. Building Sentiment Lexicon

Once sample documents were categorized by human coders and saved into the train set in data store, the Lexicon building module generates sentiment lexicon. It consists of word, the number of occurrence in positive documents, and the number of occurrence in negative documents and probability that a document is positive opinion if it contains the word, which will be used as base resource to categorize sentiment of documents in the Sentiment categorization module.

The process of building sentiment lexicon works as follows. First, it reads a document in the train set. Then it parses the document by word and checks the labeled sentiment and weight. In the comments on YouTube, a user can add a *like* or *dislike* tag, indicating the degree of user's agreement on the comments. We use the tags as a weight point. The number of occurrence for every word in positive and negative documents are counted and saved into sentiment lexicon. Finally, the probability that the document is positive opinion if it includes the word is computed for every word and saved into sentiment lexicon. Figure 22 shows the overall process and example of building sentiment lexicon using the labeled sample train set. Assume there are three documents in a train set and each document is labeled as shown in figure 11. If the word *like* appears in a document labeled positive opinion, the number of occurrence in positive opinion for the word is increased by one. If the labeled document has a *like* tag, the number of occurrence in positive opinion for the word is increased by two. If the word *like* appear in positive opinion twice and negative opinion once, the probability that a document is positive opinion will be 0.67 if the document includes the word *like*.



Labeled Sample Train Set

Documents	Sentiment	Weight (like)
I like the commercial	Positive	1
Well done	Positive	0
I don’t like godaddy ads It’s gross	Negative	0
…	…	

Parsing message by word and calculate the number of occurrence in positive and negative message

I	**like**	**the**	**commercial**				Positive	2
well	**done**						Positive	1
I	**dont**	**like**	**godaddy**	**ads**	**Its**	**gross**	negative	1

Sentiment Lexicon

Word	# of occurrence in Positive msg	# of occurrence in Negative msg	probability
like	2	1	0.67
well	1	0	1.0
gross	0	1	0.0
…	…	…	…

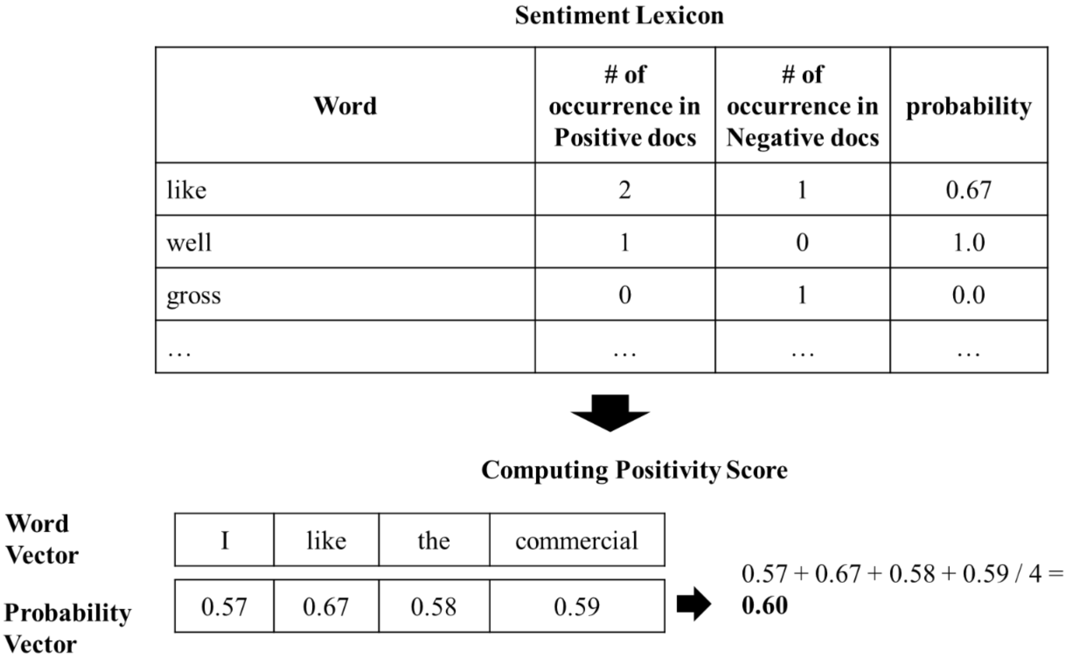
Figure 22. Example of Building Sentiment Lexicon using Labeled Sample Train Set

5.3. Categorize Comments

Once sentiment lexicon is built completely, Sentiment categorization module classifies a document into a positive or negative opinion. The document sentence is represented with vector space model (VSM) where each word in the document and its probability in sentiment lexicon are shown together. Then, the positivity score of a document is computed as follows.

$$\text{Positivity Score (d)} = \frac{\sum_{i=1}^{n} P(w_i)}{n} \qquad (8)$$

In (8), *w* is each word in a document *d* and *n* is the number of words in the document. *P* is probability of the word which is saved in sentiment lexicon with the word. Example of computing positivity score for a comment is visualized in Figure 23.



Sentiment Lexicon

Word	# of occurrence in Positive docs	# of occurrence in Negative docs	probability
like	2	1	0.67
well	1	0	1.0
gross	0	1	0.0
…	…	…	…

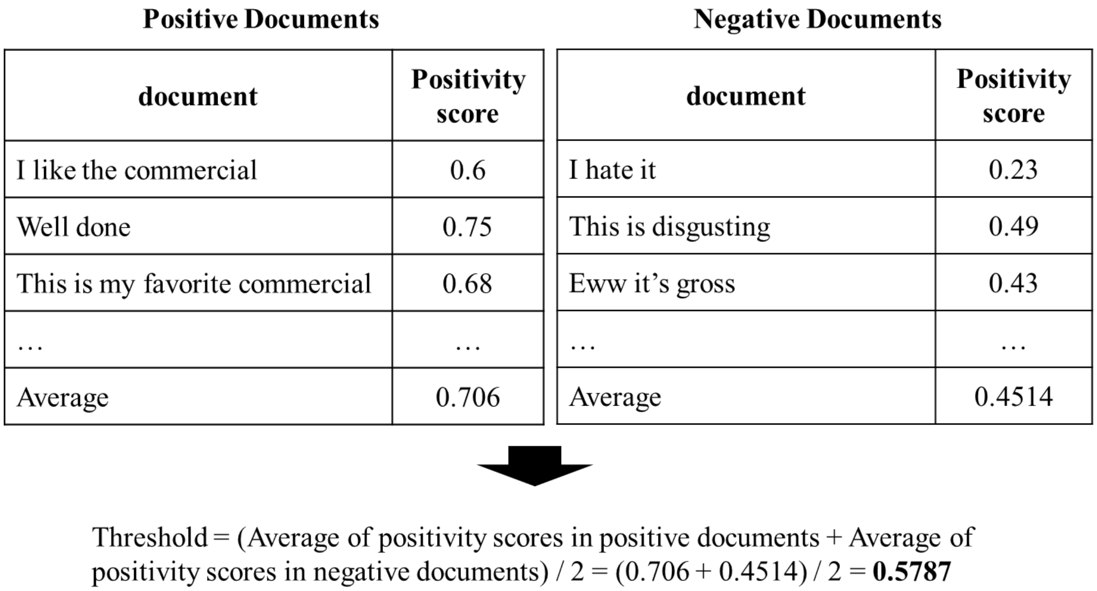
Computing Positivity Score

Word Vector	I	like	the	commercial
Probability Vector	0.57	0.67	0.58	0.59

0.57 + 0.67 + 0.58 + 0.59 / 4 = **0.60**

Figure 23. Example of computing positivity score

Once the positivity scores of all documents in train set are computed, Sentiment categorization module reads them again and computes the threshold of positivity score to classify the comment as either a positive or negative opinion. The threshold value is derived by computing mean value of positivity scores for all positive and negative documents in the train set. The example of computing threshold value is depicted in Figure 24.



Positive Documents

document	Positivity score
I like the commercial	0.6
Well done	0.75
This is my favorite commercial	0.68
…	…
Average	0.706

Negative Documents

document	Positivity score
I hate it	0.23
This is disgusting	0.49
Eww it's gross	0.43
…	…
Average	0.4514

Threshold = (Average of positivity scores in positive documents + Average of positivity scores in negative documents) / 2 = (0.706 + 0.4514) / 2 = **0.5787**

Figure 24. Computing Threshold using Positivity Scores of Positive and Negative Documents

The last step of sentiment analysis is to categorize documents in the test set using the threshold. The positivity score of each comment in the test set is computed in the same way as the previous step in the Sentiment categorization module. Then, it classifies the comment as either a positive or negative opinion. If the positivity score is greater than the threshold, it is categorized as a positive opinion. Similarly, if the positivity score is less than the threshold, it is categorized as a negative opinion. The example of classifying sentiment of comments is visualized in Figure 25.

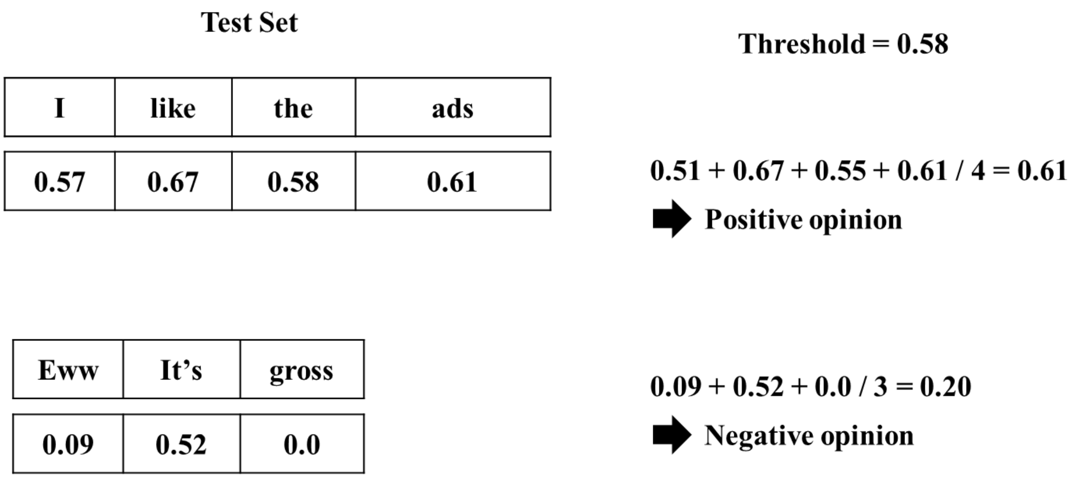


Figure 25. Example of Classifying Sentiment of Comments

Suppose a document “I like the ads” is given as shown in the Figure 5. Each word in the document is represented with VSM and the probabilities are assigned to each word (I:0.57, like:0.67, the:0.58 and ads:0.61). Then, the positivity score is computed according to the (8) and compared with the threshold value. Since the positivity score 0.61 is greater than 0.58, the document is classified as positive opinion. In the similar way, the positivity score of the second document “Eww it’s gross” is computed, compared with the threshold, and classified as negative opinion.

5.4. Sentiment Analyzer Experiment Results

Table 14 shows data collection results. We collected the video information and comments posted under the video on May 26, 2014. We collected a total of 25,003 comments for the videos. For each video, 1,000 comments are selected and used for building the sentiment lexicon and pre-processing the train data as described in the previous sections.

Table 14. Data Collection Results

Video Title	Comments Count
Official Ram Trucks Super Bowl Commercial "Farmer"	16683
Audi 2013 Big Game Commercial - "Prom"	2977
Go Daddy Bar Refaeli Kiss Super Bowl Commercial 2013 - FULL	5343

Table 15 is part of sentiment lexicon. Every word appeared in the comments is saved in the first column of sentiment lexicon. The number of word occurrence in positive and negative documents is recorded in the second and third column with the words. The probability that a document is positive opinion if it contains the word is computed using the words occurrence in positive and negative documents and is saved in the last column. As a result, sentiment lexicon was built with total of 739 words with the probability.

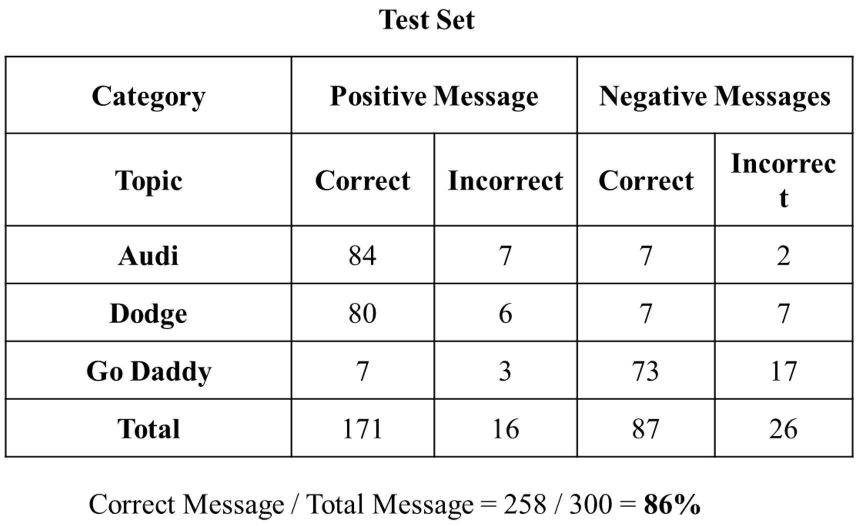
Table 15. Sentiment Lexicon

Word	The number of occurrence in positive message	The number of occurrence in negative message	Probability
love	41	0	1
great	35	5	0.87
car	23	4	0.85
pretty	9	2	0.81
all	33	8	0.8
good	17	6	0.73
dad	8	3	0.72
prom	13	5	0.72
my	50	34	0.59
make	11	9	0.55
me	25	22	0.53
not	26	29	0.47
stupid	5	6	0.45
never	6	8	0.42
kiss	5	9	0.35
why	4	8	0.33
fuck	2	10	0.16
disgusting	1	13	0.07
awkward	1	13	0.07
gross	0	20	0

Table 16. Results of Sentiment Analysis for Comments

Text(Comment)	Positivity Score	Sentiment by the proposed method	Sentiment by human coders	Results
This was the best commercial! It was so powerful........	0.67	Positive	Positive	Correct
Whoever at Dodge decided to go with this ad is a Goddamn genius!	0.64	Positive	Positive	Correct
love love love!!!!!!	1.00	Positive	Positive	Correct
Just so touching and I loved this.	0.70	Positive	Positive	Correct
Ok so I think that just made me cry a little bit. That was beautiful	0.69	Positive	Positive	Correct
VERY uncomfortable and retarded	0.41	Negative	Negative	Correct
The sound effects though.. oh goshh eww (/.)	0.36	Negative	Negative	Correct
No. I hate it	0.47	Negative	Negative	Correct
AH!! MY EYES	0.59	Positive	Negative	Incorrect
This is DISGUSTING!	0.49	Negative	Negative	Correct

To show the accuracy of the proposed algorithm, we labeled a test set in the same way as the train set is built. Then, the sentiments of documents derived by the proposed method are compared with the sentiments labeled by human coders as shown in Table 16. If human coders and the proposed method categorized a document into the same sentiment, the result is classified as correct. Otherwise, the result is classified as incorrect. The accuracy of the proposed method is computed as shown in the Figure 26. It shows that the accuracy of the proposed method is at 86%. However, the accuracy for the negative documents is relatively lower than the accuracy for positive documents, which needs to be considered and improved in future research.



Test Set

Category	Positive Message		Negative Messages	
Topic	**Correct**	**Incorrect**	**Correct**	**Incorrec t**
Audi	84	7	7	2
Dodge	80	6	7	7
Go Daddy	7	3	73	17
Total	171	16	87	26

Correct Message / Total Message = 258 / 300 = **86%**

Figure 26. Sentiment Analysis Results and Accuracy of the Proposed Method

To compare performance of the proposed method with other approaches, we applied F-measure that can be used to compute test's accuracy [18]. F-measure uses two measurement degrees; precision *p* and recall *r*. *P* is the number of correct results divided by the number of all returned results. *R* is the number of correct results divided by the number of results. The F1 score is calculated as shown in (9).

$$\mathrm{F}_1 = 2 * \frac{precision * recall}{precision + recall} \tag{9}$$

Table 17 shows results of F-measures. F1 score of our approach is 0.890 which is relatively higher than other approaches. However, it is lower than F1 score of Emoticons and SentiStrengh. Improving the accuracy needs to be considered in the future research.

Table 17. Comparison of F-Score Results

Method	**F1 score**
PANAS-t	0.737
Emoticons	0.948
SASA	0.754
SenticNet	0.810
SentiWordNet	0.789
SentiStrength	0.894
Happiness Index	0.821
LIWC	0.731
Proposed Approach	0.890

6. Dynamic Seed Selection

In the previous research [53], we presented a Java-based Twitter data crawler. It starts the data collection process from initial seed nodes, which are essentially user accounts. The initial seeds serve as the starting point of the data collection. Upon selecting seed nodes, it collects tweets the seed nodes wrote, the profiles of seed nodes and the followers of the seed nodes. Then, it collects the tweets the followers wrote, the profiles of followers, and followers of the followers, and so on. It repeats this iteration process until it reaches a pre-defined search depth or it crawls the predefined number of tweets.

Even though the crawler was able to collect a huge amount of Twitter data for specific topics, the initial seed nodes used to be selected by humans. The manual selection of seed nodes would severely damage the quality of crawled tweets. If the initial seed nodes were not properly selected, a lot of noisy data would be gathered, which makes data analysis difficult. Moreover, it could collect irrelevant tweets in the middle of the collection process too, since the crawler never evaluates seed nodes during the entire collection process.

6.1. Seed Selection Algorithm

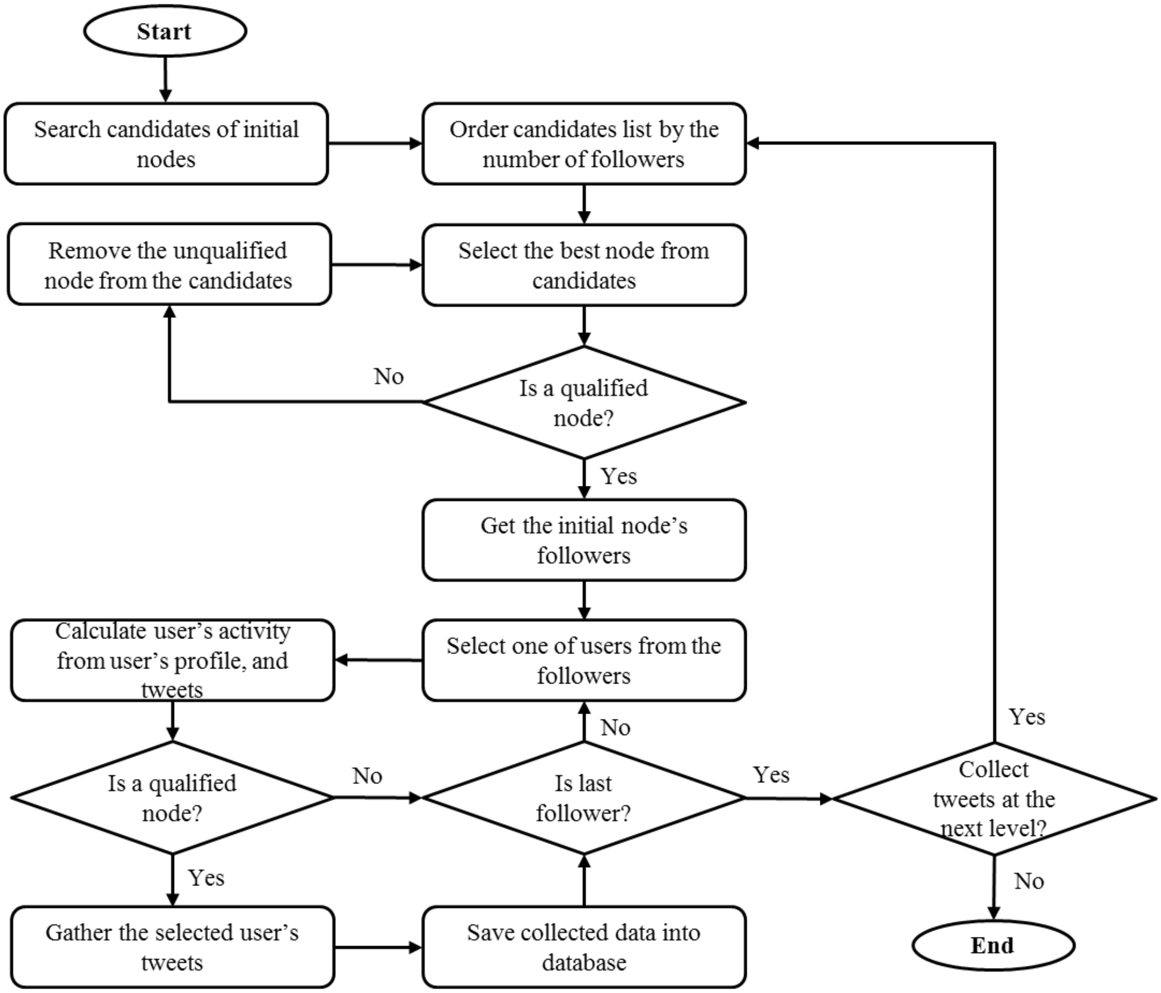

Start
Search candidates of initial nodes
Order candidates list by the number of followers
Remove the unqualified node from the candidates
Select the best node from candidates
No
Is a qualified node?
Yes
Get the initial node's followers
Calculate user's activity from user's profile, and tweets
Select one of users from the followers
Yes
No
Is a qualified node?
No
Is last follower?
Yes
Collect tweets at the next level?
Yes
No
Gather the selected user's tweets
Save collected data into database
End

Figure 27. Flow chart of data collection process

Figure 27 shows the overall data gathering process of the Twitter Crawler with the Seed handler. First, the Crawler takes a keyword as parameter. Once the Crawler is initiated, it searches for tweets that contain the keyword by use of the search() method in Twitter Java API. The search() method returns the most recent 15 tweets that contain the keyword. The 15 user accounts who wrote the tweets become the candidates for an initial node and are saved in a list. When the list of candidates is built, the algorithm calculates the activity weight of each candidate, and the list is sorted by the number of followers of each candidate. The first node that has the biggest number of followers in the list and also is a qualified node will be the initial node. If the first node in the list is not qualified, it checks the next available node to see if the node is qualified. Rebuilding a new list of candidates is needed if all candidates are not qualified. Once an initial node is selected, the crawler collects all followers of the initial node and get various information on initial node's followers such as each follower's unique id, language, number of followers, number of friends, etc. The activity weight of each follower is calculated.

If a follower is qualified, the tool collects all tweets the follower wrote and its followers will be saved for the next level search. It iterates this process until there is no more follower in the list.

The profile of a Twitter user has several properties that include the number of followers, the number of friends, the number of keyword-related tweets, the date tweeted, and the favorite count. Among them, we use the number of followers, the number of keyword-related tweets, and date tweeted as the main factors for calculating user's activity. Figure 28 shows the algorithm of calculating user's activity.

Notation: T is a set of tweets that are posted by a user over the last 30 days. K is a string variable containing a keyword. W represents the user′s activity weight. M indicates the number of tweets, and N represents the number of tweets containing the keyword W.

```
M ← T.size;
N ← 0;
for I from 0 to M
        if the tweet T[I] contains the keyword K then
                N ←  N + 1;
endfor
if M is not 0 then
        W ← N/M;
```

Figure 28. Algorithm of calculating user's activity weight

We believe that organizing user nodes by each node's activity weight allows us to discover the seed nodes that have the significant influence on the keyword. Also note that during the data collection process, the ordered list of the nodes dynamically changes due to the dynamic nature of the seed selection algorithm.

6.2. Seed Selector Experiment Results

Using the real Twitter data, we have conducted a performance evaluation of the presented algorithms. The experiments took place from Dec. 2013 to March 2014 in the United States of America. The objectives of the experiments are two-fold: to empirically determine the good value of the activity weight, and to compare the effectiveness of our algorithm in comparison with the manual selection of seeds.

The activity weight of a user represents how often the user writes tweets relevant to the given keyword. Deciding the activity weight is important for efficient collection of relevant tweets. If the activity weight is too low, the number of collected tweets tends to increase, but the number of relevant tweets tends to be low (i.e. low precision). If the activity weight is too high, the number of collected tweets tends to decrease, but the number of relevant tweets tends to be high (i.e. high precision).

We conducted an experiment to empirically derive what value of the activity weight would work out well in terms of collecting relevant tweets. The president of USA, Obama, has been chosen as a keyword to build a dataset for analysis. In this experiment we crawled the Twitter data, changing the activity weight from 0.0 to 0.6. increasing by 0.1. Figure 29 shows the experiment result. There are two bars in each column. The left bar indicates the number of all collected tweets, and the right bar indicates the number of relevant tweets. The number of all collected tweets significantly falls down as the activity weight increases. The number of relevant tweets reaches the peak point with the activity weight 0.2, though the activity weight 0.1 performs similarly. For the rest of experiments, we maintained the activity weight to be 0.2.

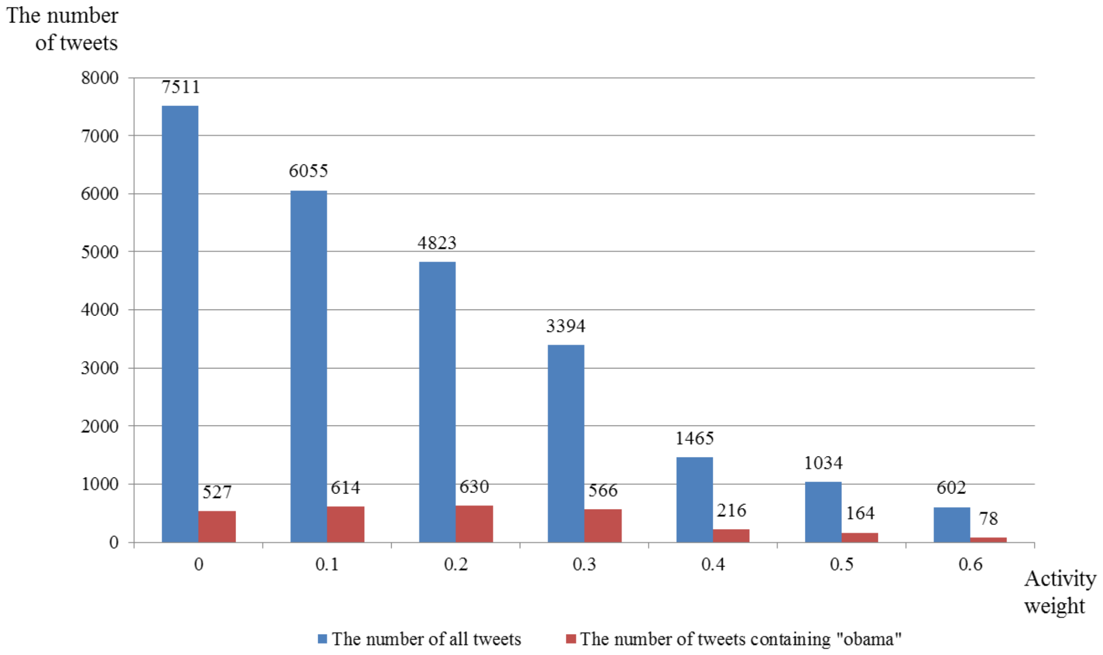


Figure 29. Activity Weight versus Crawling Effectiveness

For the second objective, two different types of data gathering approaches were used. One approach is to use the seed selection algorithm we propose, and another one is to use initial nodes manually selected by human specialists. We determined a tweet to be relevant if the tweet includes the given keyword. If not, it is treated as irrelevant.

For the former approach, we searched for tweet accounts (nodes) that posted tweets containing the given keyword in the last 30 days by Twitter API. The returned nodes were saved in a list. For saved nodes, we calculated the activity weight of each node by use of the algorithm at Figure 3. Then we sorted the nodes by the number of followers in a descending order, and checked if the first node is qualified (here the threshold was 0.2). Once we selected the initial seed node, we collected all the followers of initial seed nodes and calculated the activity weight of all followers. If the activity weight of a follower is greater than the predefined threshold (0.2 in our experiment), we collected the tweets the follower wrote. We repeated this process until there is no more follower in the list. For the latter approach we manually selected the tweet account as the initial seed node and started the collection process with the seed node. In our experiment, the depth of crawling was set to two for fast crawling.

Table 18. Effectiveness comparison of seed selection methods

Keyword		Barack Obama	Masters 2014	same sex marriage	healthcare
Manual seed selection	All tweets	2119	3503	1483	1483
	keyword related tweets	5	17	5	3
Our seed selection algorithm	All tweets	7897	4924	8015	7486
	keyword related tweets	801	279	1248	228

We conducted this experiment four times, changing the keyword each time.   The keywords we chosen are "Barack Obama", "Masters 2014", "same sex marriage" and "healthcare" that were frequently popped up in media at the time of this writing. We totally crawled eight sets of tweets (four keywords and two seed selection methods), and analyzed them. Table 18 shows the results. The "all tweets" field indicates the number of all tweets collected by an approach. The "keyword related tweets" field indicates the number of tweets that contain the topic related keyword. In overall, the number of related tweets collected by the seed analysis algorithm is much greater than that of the manual seed selection approach in all cases.

Figure 30 shows the percentages of keyword related tweets in all cases. The percentages of keyword related tweets are 0.24%, 0.49%, 0.34% and 0.2% respectively when the seed nodes are manually selected by specialists. However, the percentages of keyword related tweets are increased to 10.14%, 5.67%, 15.57% and 3.05% respectively, when the tweets are collected with the seed selection algorithm we proposed. In general, the average percentage of keyword related tweets increases from 0.31% to 8.61% with the seed selection algorithm. Those results clearly show that the seed analysis algorithm outperforms the manual seed selection in terms of collecting related tweets.

Our experiment results also show that the current method works better than the previous method [54] in terms of the effectiveness of subject-oriented tweet crawling. The current method exhibits even twice being effective for some keywords. We believe that the proper value of the activity weight could be mainly attributed to such enhancement.

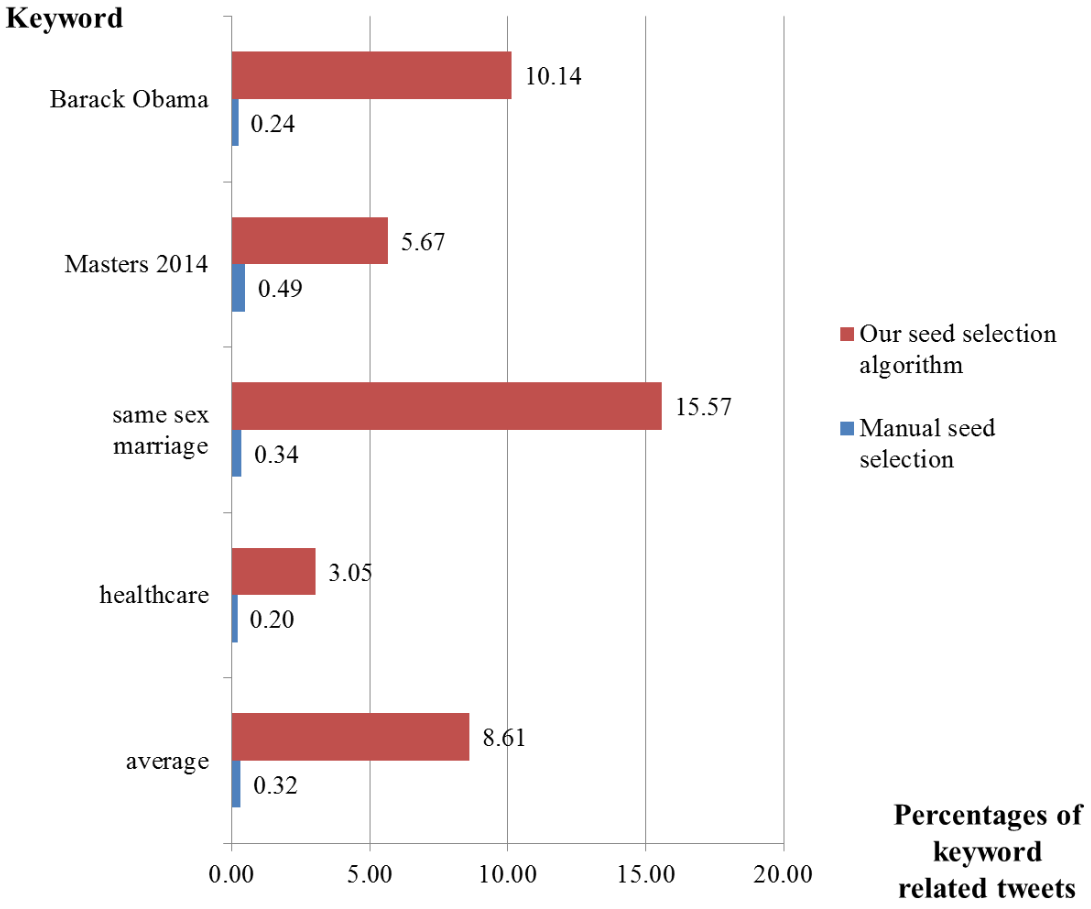


Figure 30. Graphical effectiveness comparison

7. Case Study: Data Analysis in the view of Mass Communication

7.1. A Longitudinal Analysis of Twitter Use Pattern

This study aims to address the question of how people use Twitter and the power of Twitter in creating consumer interest in brands and commercials. Another area that we are interested in this study is to examine the level of interactivity among Twitter users. In other words, whether or not people who post messages on Twitter is just posting or actually interacting with other people. Having social interaction has been identified as one of major motivations to use the media. At the same time, it was reported that the interactivity among Twitter users is lower than the expectation. Recently, many companies or web sites provide us with Twitter based advertising service and business solution. Those companies tend to generate tweets automatically for marketing. Therefore, the following research questions have been addressed by examining tweets exchanged during the Super Bowl games from 2012 to 2014.

1. What types of messages are mostly exchanged on Twitter about Super Bowl commercials? Do people posting on Twitter engage in conversation with others?
2. How do Twitter users post tweets? What kinds of devices or platforms do they prefer in using online social network?
3. Which brands have generated more traffic and buzz on Twitter? Is there a difference between private/personal websites and commercial/non-personal business sites? Do tweets from business related web sites affect the overall traffic for certain brands?

7.1.1. Methods

The study period was the three weeks over the Super Bowl game in each year; 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 release their ads on social media sites (e.g., 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, as the lingering impact of the advertising continues.

Data “Population” and “Sample” for the Analysis: Data “Population” includes all tweets exchanged from one week before the Super Bowl game day to two weeks after the game day (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, “sample” is consisted of Super Bowl commercial related tweets. A Super Bowl commercial tweet was retrieved from this “population” 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”, “Davis Beckham”, “Super Bowl” and “NFL” were used. The unit of analysis was every single tweet identified by the aforementioned search terms within the study period. Overall, the sample consisted of 3,207 tweets relating to Super Bowl commercials.

To address RQ1, we analyzed message types of data sample by year and percentages of tweets that exchanged between users. Message types of a tweet vary depending on the how a user posts the tweet on Twitter. By following the typology suggested by [55][56], we classified each tweet in data sample as three categories; Singleton, Retweet, and Reply. A Singleton is classified as an undirected message, where no specific recipient is suggested. Hence, when a user posts a tweet without referring to other users or tweets, we classified it as Singleton. When a user sends a tweet by reposting someone else's tweet, it's called Retweet and was identified by the prefix "RT." When a user posts a tweet by referring to another user with @ sign, it is considered a Reply. A Reply tweet is different from other categories in that it sends a tweet to a specifically designated person. Thus, among these three types, a Reply is considered a high-level message exchange between users. On the other hand, a Retweet can be 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 the reason, we examined the percentages of Retweet and Reply in the data sample to analyze the degree of the message exchanges between users. The higher percentages of Reply will indicate message exchanges at a higher level among tweet users than the higher percentages of Retweet or Singleton.

To address RQ2, a tweet was coded for the type of medium used to post. Twitter provides the name of the platform which contains specific uniform resource locator (URL) information, showing how each tweet was posted. Three graduate students were involved in the URL classification process. About 99.8% of all tweets were generated from top 600 URLs. All URLs are sorted by the tweet count generated from them and saved into a list. They visited the top 600 URLs in the list as the tweets generated from the 600 URLs that account for most tweet traffic (99.8% of total tweets). Then, we categorized sample tweets into two types: mobile and desktop. Then, mobile is further classified into three categories: as a Twitter official application for a specific device (e.g., for iPhone or Android), Twitter official mobile web and other mobile application. The tweets posted through the desktop are further classified into two categories: a Twitter official web and a 3rd party web (unofficial Twitter-related websites). We counted the number of sources and tweets generated from each of them to understand the device and platform preference used in posting tweets as shown in the Table 19.

Table 19. Example of Source and Tweet Count

Source	Tweet count
<a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>	569,018
<a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	352,537
web	311,690
<a href="http://blackberry.com/twitter" rel="nofollow">Twitter for BlackBerry</a>	74,061
<a href="http://twitter.com/#!/download/ipad" rel="nofollow">Twitter for iPad</a>	53,637
<a href="https://twitter.com/download/android" rel="nofollow">Twitter for Android</a>	47,635
<a href="http://www.facebook.com/twitter" rel="nofollow">Facebook</a>	38,717
<a href="http://twitter.com/tweetbutton" rel="nofollow">Tweet Button</a>	28,119
<a href="https://mobile.twitter.com" rel="nofollow">Mobile Web (M2)</a>	24,154
<a href="http://instagram.com" rel="nofollow">Instagram</a>	22,894

To address RQ3, the sources of tweets were further examined to identify whether they are from business-related (non-personal sources for profit) sites. As Table 20 indicates, tweets from “Other mobile application” and from “a 3rd party web desktop” (a total of 600 sources) were further coded into two categories: business and non-business source (personal, individual source). Here, a business source means a web site domain that provides Twitter related business advertising or analysis services. We counted the number of tweets generated from business sources and none-business sources.

In addition, we examined the number of words appeared in each category and listed top 10 frequently used brand names that aired on the Super Bowl 2014. Then, the lists of top frequently used word by group are compared with a survey from Ad Meter to see the difference between two groups and how they are differ from the survey results.

Table 20. Categories of Tweet Source Type

Category1	Category2	Category 3	Example source
Mobile	Twitter official application for a specific device (for iPhone or Android)	-	- <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a> - <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
	Twitter official mobile web	-	- <a href="https://mobile.twitter.com" rel="nofollow">Mobile Web (M2)</a> - <a href="https://mobile.twitter.com" rel="nofollow">Mobile Web (M5)</a>
	Other mobile application	Business source	-<a href="http://apps.studiohitori.com/twitrocker" rel="nofollow">TwitRocker2 for iPad</a> <a href="http://www.tweetroapp.com" rel="nofollow">Tweetro+ for Windows 8</a>
		Other	<a href="http://instagram.com" rel="nofollow">Instagram</a> <a href="http://www.apple.com" rel="nofollow">iOS</a>
Desktop	Twitter Official web	-	- web
	3rd party web	Business source	- <a href="http://unfollowers.com" rel="nofollow">Unfollowers.me</a> -<a href="https://about.twitter.com/products/tweetdeck" rel="nofollow">TweetDeck</a>
		Other	-<a href="http://www.facebook.com/twitter" rel="nofollow">Facebook</a> -<a href="http://www.hootsuite.com" rel="nofollow">HootSuite</a>

7.1.2. Experiment Results

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 (before one week- and after two weeks of the Super Bowl game day (February 5, 2012, February 3, 2013 and February 2, 2014). Out of these tweets, we analyzed only Super Bowl commercial related tweets: a total of 35,187 tweets in 2012, 34,350 in 2013, and 3,207 in 2014.

As Table 21 indicates, the overall number of Super Bowl related tweets were smaller in 2014, and the portion of commercial related tweets in 2014 was significantly decreased to 0.17%, compared to 2.49% in 2012 and 1.65% in 2013. It seems that there are fewer discussions on Twitter about Super Bowl commercials in 2014, compare to two previous years.

Table 21. Data population and sample by year(2012~2014)

Category	2012	2013	2014
Data population	1,413,524	2,079,902	1,852,181
Data sample	35,187 (2.5%)	34,350 (1.7%)	**3,207 (0.2%)**

The first research question asked what types of messages were mostly exchanged on Twitter for commercial related tweets. Table 22 shows the number tweets by message type. Over the past three years from 2012 to 2014, the most popular message type was Singleton (accounting about 61-72%), followed by Retweet (accounting for about 17-26%), and Reply (accounting for about 2%-13%). Even if each year showed a different portion of each message type, the overall pattern was consistent. The percentages of

Retweet are consistently increased from 2012 to 2014 while the portion of Reply tweets out of total tweets analyzed has been significantly decreased. In 2012 and 2013, Reply tweets accounted for 12.41% and 10.82% of the sample tweets, respectively. .However, it was significantly decreased to 2.12% in 2014. These differences were statistically significant ($\chi2$=766.01, df = 4, p <.0001), which implies the difference of tweet type is dependent on year. People post tweets about a topic on Twitter if they are interested in it. In the sense, people's interests in Super Bowl commercial were decreased in 2014. These results suggest that people's interests in a topic is proportional to tweet exchange about a topic.

Table 22. The Number of Tweets by Message type

	2012	2013	2014
Singleton	534,990 (37.8%)	763,470 (36.7%)	778,726 (42%)
Retweet (RT)	423,138 (29.9%)	727,717 (35%)	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

($\chi2$=766.0, df = 4, p <.0001)

RQ2 intended to find tweets posting method; how users post tweets and the device or platform that they use. As Table 23 shows, Twitter users used mobile devices (66.0%) more than desktop computers (34.0%) when they post tweets. Even if there are more sources for Twitter posting in desktop computers (473 sources, 78.8%) than in mobile devices (127 sources, 21.2%), two thirds of all posted tweets were by mobile devices.

This result indicates that people prefer mobile devices to desktop devices when they post tweets.

The next row shows the number of tweets generated from each device category. 1,218,497 tweets are posted from a mobile device and 629,160 tweets are generated posted a desktop device. Although the number of sources related to a mobile device is lower than desktop, the number of tweets generated from a mobile device is considerably higher than a desktop. However, the percentage of tweets generated from mobile device in data sample is relatively lower than the percentage in data population. Likewise, the percentage of tweets generated from desktop device in data sample is relatively higher than percentage in data population. We applied chi-square test on the number of data population and sample by devices. The result shows that these differences between data population and sample by device were statistically significant ($\chi2$=713.5, df = 1, p <.0001).

Table 23. The Number of Source and Tweets by Device Type

Type	Mobile	Desktop	Total
The number of sources	127 (21.2%)	473 (78.8%)	600
The number of data population (All tweets)	1,218,497 (66%)	629,160 (34%)	1,847,657
The number of data sample (Commercial related tweets)	1,298 (43%)	1,727 (57 %)	3,025

Once a device type was identified, we examined the platforms that people used to post tweets as shown in Table 24. Among the all five source types, “Other Web Pages” category was the top category, followed by the “Other Mobile Application”, accounting for 75.8% and 18.8%, respectively. However, the majority of tweets have been generated from “Mobile Twitter App for a specific device” (60.7% of all tweets) and “Twitter Official Web” (18.6%). Even though there are many applications or web pages where Twitter functionalities are integrated, the majority of people prefer Twitter official web page or Twitter mobile application to post tweets. However, trends of tweets source in data sample are considerably different from those in data population. The percentage of tweets generated from “Mobile Twitter App for a Specific Device” category in data sample is significantly decreased when compared with the percentage in data population. Furthermore, the percentage of tweets generated from “Other Web Page Category” in data sample has risen more than twice as much as the percentage in population. We also applied the Chi-square test on the number of data population and sample by source. The result of the test demonstrates that these differences between data population and sample by source were statistically significant ($\chi2$=1754.5, df = 4, p <.0001). These results indicate other web page category highly affects the traffic of commercial related tweets rather than Mobile Twitter App for a specific device category does. There is also a chance that some commercial related tweets are posted by Twitter business related sites that automatically post tweets or re-tweets for the marketing purpose in that other web page category includes all business related sites. The effects of the Twitter business related sites on traffic of data population and sample will be investigated in the next section.

Table 24. The Number of Source and Tweets by Specific Application Type

Type	Mobile Twitter App for a specific device	Twitter Official Mobile Web	Other Mobile Application	Twitter Official Web	Other Web Pages
The number of source	9 (1.5%)	5 (0.8%)	113 (18.8%)	18 (3%)	455 (75.8%)
The number of data population (All tweets)	1,121,273 (60.7%)	35,743 (1.9%)	61,481 (3.3%)	343,999 (18.6%)	285,161 (15.4%)
The number of data sample (Commercial related tweets)	1,092 (36.2%)	20 (0.7%)	186 (6.2%)	460 (15.3%)	1,257 (41.7%)

($\chi 2$=1754.5, df = 4, p <.0001)

RQ3 asked to identify the most popular brand names that were mentioned on Twitter during the Super Bowl 2014. Also, we want to find out whether profit-oriented, private business websites were successful in generating more traffic for these popular brand names.

Table 25 shows top commercial brand name in Super Bowl 2014 from all sources. Out of 58 brands that mentioned on Twitter, Budweiser was the most mentioned brand name in tweets exchanged during the study period as a total of 873 tweets contain the keyword “Budweiser”. Next brands were “Ford,” “Coca Cola,” Microsoft, and Doritos. We compared the brand names listed in Table 25 with Super Bowl commercial rankings measured by Ad Meter as shown in Table 26. The comparison yields five overlapping brand names out of ten, such as Budweiser, Coca Cola, Microsoft, Doritos, and Pepsi.

Since Ad Meter is more based on the audience's perception of strong creative elements, we do not expect the perfect overlapping.

Table 25. Top Commercial Brand Name in Super Bowl 2014 from All Sources

Rank	Company	Tweet Count	Percentages
1	budweiser	873	28.9%
2	cocacola	683	22.6%
3	maserati	505	16.7%
4	doritos	318	10.5%
5	pepsi	193	6.4%
6	microsoft	146	4.8%
7	ford	131	4.3%
8	chrysler	121	4.0%
9	kia	91	3.0%
10	audi	62	2.0%

Table 26. Super Bowl Commercial Ranking in 2014 by AD Meter

Rank	Company	Spot	Quarter	Score
1	Budweiser	Puppy Love	4	8.29
2	Doritos	Cowboy Kid	4	7.58
3	Budweiser	Hero's Welcome	3	7.21
4	Doritos	Time Machine	1	7.13
5	Radio Shack	Phone Call	1	7
6	Hyundai	Sixth Sense	1	6.87
7	General Mills Cheerios	Gracie	1	6.75
8	Microsoft	Technology	4	6.65
9	Coca Cola	Going All the Way	4	6.42
10	Pepsi	Soundcheck	HT	6.3

When we analyzed the tweets generated only from profit-oriented, business sources interesting findings emerged as shown Table 27. First, Doritos was not mentioned at all in the list while Doritos was made in top 10 AD Meter list. Second, Pepsi was the most mentioned brand through these business sources while Budweiser was the top brand in Table 25 and Table 26. Pepsi was ranked on the 5th in Table 25 and was at the bottom in Table 26. Third, three brands names (Chrysler, CarMax, and Kia) that did not appear on Table 27 were made in the top 10 list by business sources. Overall, we found that the keyword rankings in tweets generated from business sources are different from those generated from all sources and on Ad Meter rankings. It implies that those brands in Table 27 were intentionally pushed by marketers to create social buzz on Twitter.

Table 27. Top Commercial Brand Names in Super Bowl 2014 from Twitter Related Business Sources

Rank	Company	Tweet Count	Percentages
1	pepsi	86	16.7%
2	cocacola	62	12.0%
3	budweiser	61	11.8%
4	chrysler	40	7.8%
5	audi	30	5.8%
6	maserati	29	5.6%
7	dannon oikos	26	5.0%
8	microsoft	25	4.8%
9	bank of america	21	4.1%
10	carmax	20	3.9%

In this study, we identified 84 sources that are associated with Twitter business and advertising. As shown in the Table 28, only 6.1% of tweets were generated from business sources while the majority of tweets (93.9%) were generated from non-business sources. This indicates that tweets from business sources do not affect overall traffic of Twitter and they tend to post commercial and business related tweets rather than normal tweets.

Table 28. The Number of Sources and Tweets by Business and None-Business Category

Type	Non-business	Business
The number of source	516 (86%)	84 (14%)
The number of data population (All tweets)	1,847,657 (93.9%)	119,547 (6.1%)
The number of data sample (Commercial related tweets)	2,509(82.9%)	516(17.1%)

This study aimed to address how interests in a topic affect tweets exchanges about the topic, what kinds of device people prefer to post tweets and how tweets generated from business related sources affect overall traffic of Twitter. 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.

We collected all tweets that were exchanged one week before and two weeks after Super Bowl 2012, 2013 and 2014 and classified them as “Population”. Out of this data set, the Super Bowl commercial related tweets were retrieved and classified as “Sample”. To address RQ1, we examined percentages of data sample in 2012, 2013 and 2014, and ratio of tweet exchange between users. We found that tweet traffic about Super Bowl commercials is significantly decreased in 2014. Also, ratio of direct message exchange also is significantly decreased in 2014. From the results we found that tweet exchanged between users about a topic is proportional to interests in the topic.

The second research question intended to address device or platform preference of users on Twitter. We analyzed the source information that is provided with tweet to address the question. We found that people prefer mobile device to desktop when they post tweets. We also found that people still prefer official web page and mobile application provided by Twitter even though there are lots of application or web pages available where Twitter functionalities are integrated. In addition, we found commercial related tweets posting are more affected by other web pages category than the other categories. We concluded that there is a chance that some commercial related tweets are posted by Twitter business related sites that automatically post tweets or re-tweets for the marketing purpose since other web page category includes all business related sites. Thus, we investigated the effects of Twitter business related sites on traffic of data population and sample in the last experiments.

Third research question intended to examine effects of Twitter based advertising service and business solution on overall trends and traffic of Twitter. We analyzed tweets generated from business related sources to address to question. We found that tweets from business source do not affect overall traffic of Twitter and they tend to post commercial and business related tweets rather than normal tweets. The finding also shows trends of the tweets generated from business sources are affected by the companies that intentionally advertise products or companies but they do not affect top commercial brand names and overall trends of Twitter.

7.2. Users' Spatial Twitter Usage Patterns

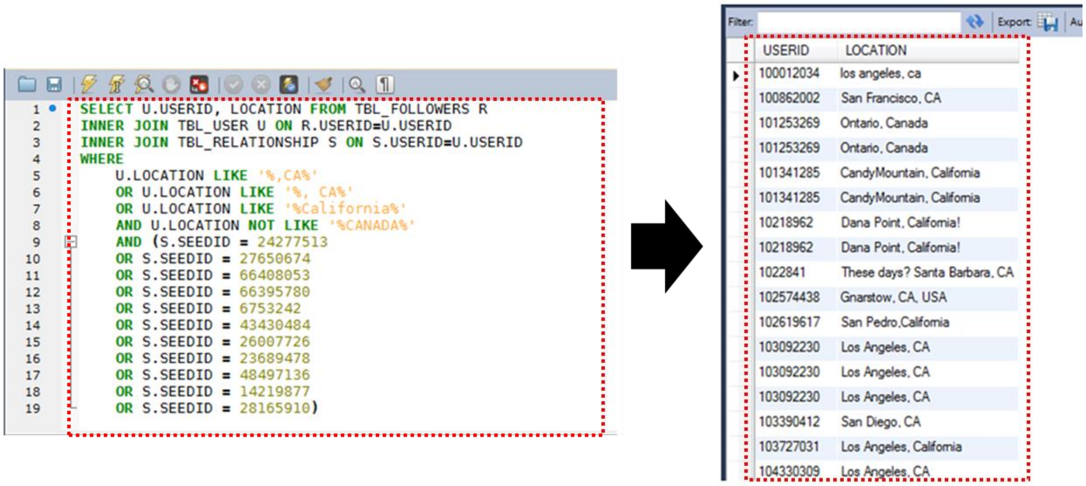
These days, Twitter message production patterns have been investigated by other research. However, the Twitter users' spatial usage patterns have not been studied in depth because of the limited information of user's profile in Twitter. For this reason, this experiment explores the Twitter users' spatial usage patterns and how these patterns are related to real world phenomenon using users' profile and following-follower relationship with the companies that broadcasted their commercials during Super Bowl 2012. The following research questions have been addressed by this experiment.

1. How people prefer car company brand? Do the brand preferences vary depending on the state and reflect real world phenomenon?
2. Are there any association rules between brand preference and regional information?

7.2.1. Methods

To address RQ1, we have collected tweets exchanged one week before and two weeks after the Super Bowl (Jan. 29, 2012 to Feb. 19, 2012). First, location and followers information were extracted from users' account that wrote the tweets. Second, we retrieved users' accounts that follow at least one car company and location field information from the user's account. The location field in Twitter accounts allows users to enter string format data. Twitter users usually fill out state information in their location field. For this reason, we filtered the location field information and classified users into 50 U.S state category. Figure 31 shows the example of filtering users' location field to find users who live in California. If the location field of a user contains ",CA", ", CA", or "California", we assumed that the user lives in California. Also, we excluded users whose location field contains “CANADA” in their location field. The number of users by each state were extracted in a similar way as users who live California were extracted. In addition, these users were classified into the 11 car company categories based on their follower information to analyze brand preference for the companies by state.

To address RQ2, we assigned additional features to state brand preference data that were used in the first experiment. First, all U.S states were divided into nine divisions (Division 1: New England, Division 2: Mid Atlantic, Division 3: East North Central, Division 4: West North Central, Division 5: South Atlantic, Division 6: East South Central, Division 7: West South Central, Division 8: Mountain, Division 9: Pacific) as shown in Figure 21. Apriori algorithm was applied to the data that were featured with the region and brand preference ranking information to find association rules.



```sql
SELECT U.USERID, LOCATION FROM TBL_FOLLOWERS R
INNER JOIN TBL_USER U ON R.USERID=U.USERID
INNER JOIN TBL_RELATIONSHIP S ON S.USERID=U.USERID
WHERE
    U.LOCATION LIKE '%,CA%'
    OR U.LOCATION LIKE '%, CA%'
    OR U.LOCATION LIKE '%California%'
    AND U.LOCATION NOT LIKE '%CANADA%'
    AND (S.SEEDID = 24277513
    OR S.SEEDID = 27650674
    OR S.SEEDID = 66408053
    OR S.SEEDID = 66395780
    OR S.SEEDID = 6753242
    OR S.SEEDID = 43430484
    OR S.SEEDID = 26007726
    OR S.SEEDID = 23689478
    OR S.SEEDID = 48497136
    OR S.SEEDID = 14219877
    OR S.SEEDID = 28165910)
```

USERID	LOCATION
100012034	los angeles, ca
100862002	San Francisco, CA
101253269	Ontario, Canada
101253269	Ontario, Canada
101341285	CandyMountain, California
101341285	CandyMountain, California
10218962	Dana Point, California!
10218962	Dana Point, California!
1022841	These days? Santa Barbara, CA
102574438	Gnarstow, CA, USA
102619617	San Pedro,California
103092230	Los Angeles, CA
103092230	Los Angeles, CA
103092230	Los Angeles, CA
103390412	San Diego, CA
103727031	Los Angeles, California
104330309	Los Angeles, CA

Figure 31. Filtering User Location Field (Find users who live in California)

7.2.2. Experiment Results

Table 29 shows the number of followers to 11 car companies that advertised their commercials in Super Bowl 2012. The number of followers of the companies was categorized by U.S states again. Overall, the number of total users in each state reflects population in it. Brand preferences of the companies in the all U.S states were visualized in Figure 32.

Table 29. Number of followers to 11 car companies in U.S state.

	Toyota	Kia	Acura	Hyundai	Audi	Volkswagen	Honda	Lexus	Chevrolet	Cadillac	Chrysler	Total
AL	59	111	63	87	53	43	58	54	101	74	73	776
AK	12	7	3	11	8	2	6	6	12	6	3	76
AZ	109	96	84	109	98	83	76	81	138	84	96	1054
AR	57	67	41	50	83	56	49	61	101	61	61	687
CA	846	660	801	779	731	610	656	580	695	569	664	7591
CO	143	99	110	132	147	107	103	95	137	101	130	1304
CT	36	42	58	32	34	38	34	33	43	34	35	419
DE	25	25	20	22	28	25	19	24	31	23	25	267
FL	293	412	300	293	321	287	236	272	386	285	312	3397
GA	152	272	117	136	134	117	104	119	195	120	145	1611
HI	16	20	16	15	16	10	14	11	16	8	10	152
ID	14	15	12	12	14	15	9	9	16	14	18	148
IL	155	173	147	162	153	146	128	143	206	173	207	1793
IN	180	143	111	158	211	166	176	179	231	165	165	1885
IA	76	94	67	62	51	59	64	48	123	72	100	816
KS	75	95	71	67	70	56	51	63	115	73	76	812
KY	46	49	39	39	37	41	35	37	70	40	41	474
LA	41	62	27	45	42	29	23	35	50	30	41	425
ME	38	27	21	24	54	43	28	27	58	33	59	412
MD	73	96	80	72	66	67	61	64	80	56	63	778
MA	147	142	148	119	151	115	138	138	161	124	149	1532
MI	356	300	190	310	314	318	243	199	583	493	1040	4346
MN	52	61	54	52	50	46	57	44	77	56	71	620
MS	12	44	9	13	14	10	11	21	33	19	11	197
…	…	…	…	…	…	…	…	…	…	…	…	…
Total	5182	5863	4833	4926	5140	4556	4284	4241	6540	4809	5980	56354

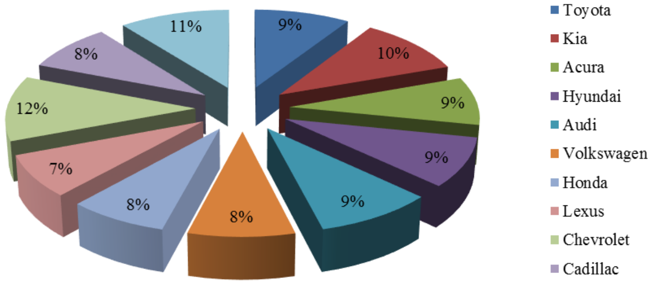


Figure 32. Brand Preference of 11 Car Companies in the U.S.

Figure 33 compares brand preferences in all U.S. states with that of Michigan, Georgia and Alabama where Chrysler, Kia and Hyundai manufactures are located. The average brand preference for Chrysler in all U.S states is 11%. However, the average brand preference for Chrysler in Michigan is higher than the average by 218% as it accounts for 24%. Also, the average brand preferences for Kia and Hyundai in Georgia and Alabama are higher than average brand preferences to the companies. Overall, the average brand preference for a car company in a state where the manufacturer of the company is located is higher than the average of brand preference in all U.S state. These results indicate that those brand preference data reflect a real world phenomenon.

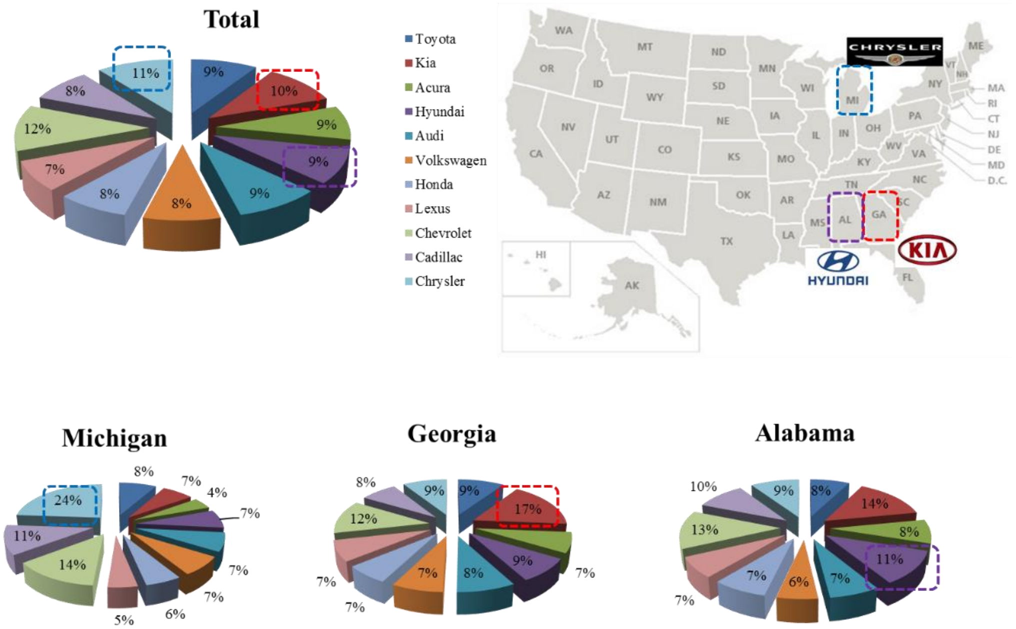


Figure 33. Comparison of Brand Preference in all U.S. State with Michigan, Georgia and Alabama

To apply association rules, region information is assigned to each state as shown in the Figure 34. Then, Apriori algorithm is applied as explained in the section 7.2.1. Following is the list of association rules, confidence of which is greater than 0.75. Overall, people who live in west north central (Division 4) prefer Chevrolet mostly. Also, people who live in south Atlantic prefer Kia.

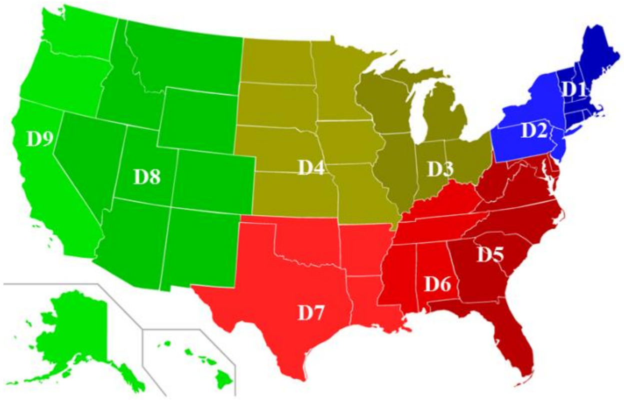


Figure 34. Region of the U.S

- There are 7 states in region D4 ==> The first brand is Chevrolet in 7 states, confidence : (1)
- There are 5 states in which the second brand is Chevrolet and region is D5, ==> The first brand is Kia in 5 states, confidence : (1)
- There are 9 states in region D5 ==> The first brand is Kia in 7 states, confidence : (0.78)

8. Conclusions

In this paper, we have proposed an integrated social networks data collection, management and analysis system, and specific analysis methods for topic classification and sentiment analysis. In order to study the broad spectrum of social networks, we have developed the Data Crawler that collects data from three source channels, Twitter, YouTube and New York Times. The Term extractor, Topic Classifier and Sentiment Analyzer are integrated into the Analyzer to pre-process documents collected from social networks. We also integrated the Hadoop into the system to manage huge volume of data efficiently.

We developed Topic classifier using several machine learning algorithms and voting systems. The Topic Classifier trains existing topic related documents using SVM, Naïve Bayes and decision tree and classifies new document using voting results of the three classifiers. We tested the accuracy of the Topic Classifier using three types of topic related documents; advertising, opinion and stock. The experiment results show 0.93 of accuracy at average. The Topic Classifier was used to select additional documents for train set. After the automatically selected documents are added into train set, the accuracy of the Topic Classifier were tested again. The experiment with new train set was also successful in that the accuracy of algorithm with new train set remained almost same as the accuracy with initial train set.

We developed a sentiment analysis algorithm using a probability model that guarantees relatively higher accuracy than existing approaches with broader application. The result shows that the proposed approach outperforms most existing sentiment analysis approaches in terms of accuracy. In addition, the proposed approach was implemented only using text information without requiring any additional information. This proposed approach, however, has a limitation that requires preprocessing of sample text documents by human coders. We will investigate a fully automated sentiment analysis method in the next research, and continue to work on improving the accuracy rate of a proposed method.

We proposed a seed account selection algorithm in order to effectively collect Twitter data relevant to given keywords. The algorithm evaluates user's activities and updates the seed nodes dynamically. In our experiments, we compared two approaches with real Twitter data, one using the proposed algorithm and one relying on human experts. Our experiments show that the proposed selection outperforms the manual selection in terms of crawling relevant tweets. The beauty of a subject-oriented crawler is to get a sufficient number of relevant tweets in a reasonable time. As a future work, we plan to work on the scale-ability of the Crawler. The Crawler should be scale-able enough to crawl a vast amount of Twitter data by adding necessary resources (such as servers). Moreover, we determined a tweet to be relevant to given keywords if the tweet has the keywords in the message. It is worth investigating "being relevant" in the future research. We believe that notable techniques (such as the vector model) commonly employed in the field of information retrieval could be adapted in Twitter crawlers. We leave this issue as a future work.

Additionally, we conducted interdisciplinary research on social networks data analysis with mass communication department. Documents about Super bowl Advertising have been examined using statistical analysis techniques. We found several interesting facts about social networks phenomenon. Many social networks analysis results coincide with real world facts. We concluded that social network analysis results can be used as reference data to predict real world phenomenon.

Contributions of this paper are summarized as follows. First, this study suggested Hadoop based data architecture and system to manage and analyze a huge amount of data generated from social networks. Second, we proposed automatic data crawler that collects documents from multiple source channels. Third, we proposed a topic classifier and a sentiment analysis algorithm based on term based features and probabilistic model, which guarantees higher accuracy than other approaches. Fourth, we proposed a dynamic seed selection algorithm that allows us to collects documents relevant to certain keywords in Twitter. Lastly, we applied data in social networks and data analysis techniques to the analysis in other domain.

References

[1] 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.

[2] 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

[3] 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

[4] 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

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] J. Tao, F. Zheng, A. Li and Y. Li, "Advances in Chinese Natural Language Processing and Language Resources", Proceedings of the Speech Database and Assessments, 2009 Oriental COCOSDA International Conference, August 2009, pp.13-18, ISBN : 978-1-4244-4400-7

[13] C. Surabhi, "Natural Language Processing Future", Proceedings of International Conference on Optical Imaging Sensor and Security, July 2013, pp.1-3, ISBN : 978-1-4799-0935-3

[14] M. Azadnia, S. Rezagholizadeh and A. Yari, "Natural Language Processing Laboratory Plan", Proceedings of 2010 6th International Conference on Networked Computing (INC), May 2010, pp.11-13, ISBN: 978-89-88678-20-6

[15] H. Wahl, W. Winiwarter and G. Quirchmayr, "Natural Language Processing Technologies for Developing a Language Learning Environment", Proceedings of the

12th International Conference on Information Integration and Web-based Applications & Services, November 2010, pp.381-388, ISBN: 978-1-4503-0421-4

[16] The Stanford Natural Language Processing Group, http://nlp.stanford.edu

[17] A. Sharma and S. Dey, “A comparative study of feature selection and machine learing techniques for sentiment anlaysis”, Proceedings of the 2012 ACM Research in Applied Computation Symposium, October 2012, pp. 1-7, ISBN:978-1-4503-1492-3.

[18] P. Goncalves, M. Araújo, F. Benevenuto, and M. Cha, “Comparing and combining sentiment analysis methods”, Proceedings of the first ACM conference on Online social networks, October 2013, pp. 27-38, ISBN: 978-1-4503-2084-9.

[19] P. Melville, W. Gryc, and R. D. Lawrence, “Sentiment analysis of blogs by combining lexical knowledge with text classification”, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, June 2009, pp. 1275-1284, ISBN: 978-1-60558-495-9.

[20] B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From Tweets to Polls:Linking Text Sentiment to Public Opinion Time Series”, Proceedings of the International AAAI Conference on Weblogs and Social Media, May 2010, pp. 122-129.

[21] T. Winson et al., “OpinionFinder:A System for Subjectivity Analysis”, Proceedings of HLT/EMNLP 2005 Interactive Demonstrations, October 2005, pp. 34-35, doi:10.3155/1225733.1225751.

[22] Y. He, “A Bayesian Modeling Approach to Multi-Demensional Sentiment Distributions Predictions”, Proceedings of the Frist International Workshop on Issues

of Sentiment Discovery and Opinion Mining, August 2012, Article No.1, ISBN:978-1-4503-1543-2.

[23] A. Sharma and S. Dey, “A Boosted SVM based Sentiment Analysis Approache for Online Opinionated Text”, Proceedings of the 2013 Research in Adaptive and Convergent Systems, October 2013, pp. 28-34, ISBN: 978-1-4503-2348-2.

[24] A. Sharma and S. Dey, “A comparative study of feature selection and machine learning techniques for sentiment analysis”, Proceedings of the 2012 Research in Adaptive and Convergent Systems, October 2012, pp. 1-7, ISBN:978-1-4503-1492-3.

[25] P. H. Guerra, A. Veloso, W. Meira, and V. Almeida, “From Bias to Opinion: A Transfer-Learning Apporach to Real-Time Sentiment Analysis”, Proceddings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining(KDD 11), August 2011, pp. 150-158, ISBN:978-7-4503-0813-7.

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

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

[28] X. Yan, J. Guo, Y, Lan and X. Cheng, “A biterm topic model for short texts”, Proceedings of the 22nd international conference on World Wide Web, May 2013, pp.1445-1456, ISBN:978-1-4503-2035-1.

[29] Q. Mei, X. Shen and C. Zhai, “Automatic labeling of multinomial topic models”, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, August 2007, pp.490-499, ISBN:978-1-59593-609-7

[30] M. Rosen-Zvi, C. Chemudugunta, T. Griffiths. P. Smyth and M. Steyver, “Learning authoer-topic models from text corpora”, ACM Transactions on Information Systems, Volume 28, Issue1, January 2010.

[31] J. Paik, “A novel TF-IDF weighting scheme for effective ranking”, Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, July 2013,pp.343-352, ISBN:978-1-4503-2034-4

[32] T. Roelleke and J. Wang, “TF-IDF Uncovered: A Study of theories and Probabilites”, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, July 2008, pp.435-442, ISBN: 978-1-60558-164-4

[33] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet Allocation”, Journal of Machine Learning Research, Vol.3, March 2003, pp.993-1022.

[34] C. Cortes and V. Vapnik, “Support Vector Networks”, Machine Learning, Vol.20, November 1995, pp.273-297.

[35] T. Joachims, “Text Categorization with Support Vector Machines: Learning with Many Relevant Fetures”, Proceedings of the 10th European Conference on Machine Learning, 1998, pp.137-142, ISBN:3-540-64417-2

[36] N. Friedman, D. Geiger and M. Goldszmidt, “Bayesian Network Classifiers”, Machine Learning, vol. 29, pp 131-163, 1997

[37] R. Irina, “An empirical study of the naïve Bayes classifier”, IJCAI Workshop on empirical Methods in AI, 2001

[38] L. Jiang, H. Zhang and Z. Cai, “A Novel Bayes Model: Hidden Naïve Bayes”, IEEE Transactions on Knowledge and Data Engineering, Vol.21, No.10, October 2009.

[39] J.R. Quinlan, “Induction of Decision Trees”, Machine Learning, vol.1, pp.81-106, 1986.

[40] F. Esposito, D. Malerba and G.Semeraro, “A Comparative Analysis of Methods for Pruning Decision Trees”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5, pp.476-491, May 1997.

[41] C. Z. Janikow, “Fuzzy Decision Trees : Issues and Methods”, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 28, No. 1, pp.1-14, February 1998.

[42] J. R. Quinlan, “C4.5: Programs for Machine Learning”, San Mateo, Calif.:Morgan Kaufmann, 1993.

[43] L. Breiman, J. Friedman, C. J. Stone and R. A. Oslhen, “Classification and Regression Trees”, Chapman and Hall/CRC, 1984

[44] J. Han, M. Kamber and J, Pei, “Data Mining Concepts and Techniuqes”, 3rd edi, Morgan KaufMann, 2011

[45] M. Arias, A. Arratia and R. Xuriguera, “Forecasting with Twitter Data”, ACM Transactions on Intelligent Systems and Technlogy (TIST), Vol. 5, Issue 1, December 2013.

[46] G. Mishne and N. Glance, “Predicting Movie Sales from Blogger Sentiment”, In Proceedings of the AAAI Symposium on Computational Approaches to Analysing Weblogs, 2006, pp.155-158.

[47] S. Asur and B. A. Huberman, “Predicting the future with social media”, In Proceedings of the IEEE/WIC/ACM Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010, pp.492-499, ISBN:978-0-7695-4191-4

[48] J. Bollen, H. Mao and X. Zeng, “Twitter mood predicts the stock market”, Journal of Computational Science, Vol 2, Issue 1, March 2011, pp.1-8.

[49] M. S. A. Wolfram, “Modelling the stock market using twitter”, M.S. thesis, Schools of Informatics, University of Edinburgh, 2010.

[50] S. Wakamiya, R. Lee and K. Sumiya, “Crowd-powered TV viewing rates: measuring relevancy between tweets and TV programs”, Proceedings of the 16th international conference on Database systems for advanced applications, pp.390-401, 2011, ISBN:978-3-642-20243-8

[51] V. Lampos, T. D. Bie and N. Cristianini, “Flu detector - Tracking epidemics on twitter”, In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases, pp.599-602, 2010

[52] H. Lee, Y. Han, Y, Kim and K, Kim, “Sentiment Analysis on Online Social Network Using Probability Model”, In Proceedings of the Sixth International Conference on Advances in Future Internet, pp.14-19, 2014

[53] C. Byun, H. Lee, Y. Kim, and K. Kim, “Automated Twitter Data Collecting Tool and Case Study with Rule-based Analysis,” In Proceedings of 14th International

Conference on Information Integration and Web-based Application & Services (IIWAS), pp. 196-204, 2012.

[54] C. Byun, H. Lee, J. You, and Y. Kim, "Efficient Keyword-related Data Collection in a Social Network with Weighted Seed Selection," International Journal of Networked and Distributed Computing, Vol. 1, No. 3, pages 167-173, August 2013.

[55] H. Kwak, C. Lee, H. Park and S. Moon, “What is Twitter, a social network or a news media?”, Paper presented at the 19th International conference on WWW, Raleigh, NC, April 2010.

[56] A. O. Larsson, “Twitting the viewer –Use of Twitter in a talk show context”, Journal of Broadcasting & Electronic media 57 (2), pp. 135-152, 2013

[57] I. King, J. Li, and K. T. Chan, “A brief survey of computational approaches in social computing”. In IJCNN’09: Proceedings of the 2009 international joint conference on Neural Networks, pp. 2699–2706, Piscataway, NJ, USA, 2009. IEEE Press. ISBN:978-1-4244-3549-4

[58] C. Byun, H. Lee, J. You and Y. Kim, “Dynamic Seed Analysis in a Social Network for Maximizing Efficiency of Data Collection”, Proceedings of the Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2013 14th ACIS International Conference, July 2013, pp. 132-136

[59] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” Proceedings of the 19th International Conference on World Wide Web, pages 851-560, Raleigh, 2010.

[60] E. Aramaki, S. Maskawa, and M. Morita, “Twitter Catches The Flu: Detecting Influenza Epidemics using Twitter,” Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1568-1576, Edinburgh, 2011.

[61] D. Correa and A. Sureka, “Mining Tweets for Tag Recommendation on Social Media,” Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents, pages 69-76, Glasgow, 2011.

[62] P. Noordhuis, M. Heijkoop, and A. Lazovik, "Mining Twitter in the Cloud," Proceedings of the IEEE 3rd International Conference on Cloud Computing, 2010.

Curriculum Vita

Personal Information

Hyeonchoel Lee

Department of Computer and Information Sciences
Towson University
8000 York Road
Towson, Maryland 21252
USA

[redacted]

Formal Education

2011	Master of Science, Computer and Information Science Towson University, Towson, MD
2006	Bachelor of Engineering, Computer System Hansung University, Seoul, South Korea