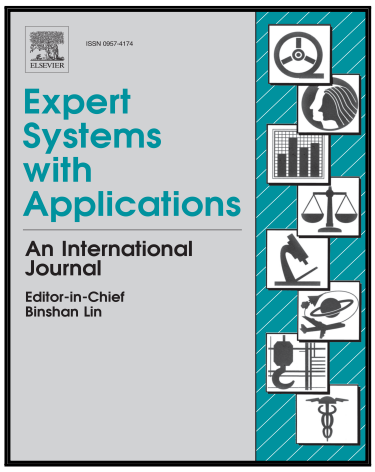


Social Media Contents based Sentiment Analysis and Prediction System

SoYeop Yoo , JeIn Song , OkRan Jeong

PII: S0957-4174(18)30212-4
DOI: 10.1016/j.eswa.2018.03.055
Reference: ESWA 11902

To appear in: *Expert Systems With Applications*

Received date: 5 July 2017
Revised date: 2 March 2018
Accepted date: 26 March 2018

Please cite this article as: SoYeop Yoo , JeIn Song , OkRan Jeong , Social Media Contents based Sentiment Analysis and Prediction System, *Expert Systems With Applications* (2018), doi: 10.1016/j.eswa.2018.03.055

Highlights

- We proposed social media contents based sentiment analysis and prediction system.
- The system detects events in real time out of the massive social media contents.
- Users' sentiments are classified using the Convolutional Neural Networks.
- The next sentimental path is predicted through Long Short-Term Memory.

Title
Social Media Contents based Sentiment Analysis and Prediction System

Author
SoYeop Yoo, Gachon University , 1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Korea.
bbusso90@gmail.com

JeIn Song[1], Gachon University , 1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Korea.
wpdls601@gc.gachon.ac.kr

Corresponding author
OkRan Jeong, Gachon University, 1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Korea.
(+82)-31-750-5831
orjeong@gachon.ac.kr

[1] Present address: ZUM Internet Corp., Banpo-daero 3, Seocho-gu, Seoul, Korea.

Abstract
With the influence and social ripple effect of social media sites, diverse studies are in progress to analyze the contents generated by users. Numerous contents generated in real time contain information about social issues and events such as natural disasters. In particular, users show not only information about the events that occurred but also their sentiments. In this paper, we propose a system for analyzing and predicting users' sentimental trajectories for events analyzed in real time out of the massive social media contents, and show the results of preliminary validation work that we have done. We show both trajectory analysis and sentiment analysis so that users can obtain the insight at a glance. Also, we increased the accuracy in sentiment analysis and prediction by making use of the latest deep-learning technique.

Keywords
social media; sentiment analysis; sentiment prediction; sentimental trajectory

Acknowledgements
This research was supported by Basic Science Research Program through the NRF(National Research Foundation of Korea), and the MIFP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW(2015-0-00932) supervised by the IITP(Institute for Information & communications Technology Promotion (Nos.NRF 2015R1C1A2A01051729, 2015-0-00932).

1. Introduction
As the usage of social media sites increased, sites such as Twitter and Instagram has become to have various meanings rather than simply remaining as site on which contents are prepared and uploaded. Because of the characteristics of social media sites that can be accessed always without any temporal or spatial restriction and can be connected easily and quickly if only the Internet has been connected, users unceasingly create small and large contents ranging from their mere trifles to social issues and disasters. In recent years, the influence of social media sites on everyday life have become so large that even information on large and small incidents or disasters is obtained through social media sites (Hu, Jamali, & Ester, 2012; YD Kim & Moon, 2011; Lawton, 2001).

Due to the influence and social ripple effect of social media sites, diverse studies are in progress to analyze the contents generated online. Social media contents analyses are conducted in diverse methods and for diverse purposes. Among numerous contents, especially those texts that are firsthand written by users contain the most direct and important information. Since the contents are created according to the users' intentions, the time of creation time also becomes an important factor in contents analysis (Hu et al., 2012; YD Kim & Moon, 2011; Lawton, 2001; Tang, Chang, & Liu, 2013; Zafarani, Abbasi, & Liu, 2014).

Among social media sites, Twitter has very high communicability of information on social issues, disasters, and accidents in particular thanks to its characteristics that information is delivered with short letters not exceeding 140 characters and that everybody can easily get and propagate information. If Twitter is analyzed focusing on events such as social issues and disasters, events that have occurred can be identified and the trajectories of the identified events such as the propagation paths can be analyzed. In addition, if the sentiments of peoples in various regions on the relevant event are analyzed, many pieces of information on a certain event can be obtained. Sentimental path analyses enable sentiment analyses according to places and time and the prediction of users' sentiments in advance (Cho, Myers, & Leskovec, 2011; Feng & Zhu, 2016; Scellato, Noulas, Lambiotte, & Mascolo, 2011; Senaratne, Bröring, & Schreck, 2014; Zheng & Zhou, 2011).

To analyze the sentimental trajectory of the keywords reflected in contents, a database system that efficiently stores and manages these social media contents is necessary. Social media contents are characterized by the fact that not only they are semi-structured data, but also they create massive data in real-time. As a database for structural and quantitative management of social media contents, in this study, AsterixDB, which has been developed as an Apache project and open source, is used. Since it has a structure flexible to semi-structured data and provides diverse indexing methods, it has characteristics that facilitate social media contents analysis

(Alsubaiee, Altowim, Altwaijry, & Behm, 2014).
Social media contents spread swiftly while users communicate and share them freely within the site. In particular, when an event such as a social issue or a natural disaster occurs, a numerous data related to the relevant keyword are poured out so that the fact that the particular event has occurred can be identified by analyzing the data. The sentiment can be also predicted by analyzing the sentimental path utilizing the region and temporal elements of the extracted event in real time.

In this paper, our major contributions are as follows.
1) **Efficiency on cost**; We use AsterixDB to handle social media contents efficiently and enable to auto classification without directly labeling positive/negative. To build a sentiment classification model, labeled data are needed and usually many research use data that is labeled directly by people. Instead of aforementioned method, we collect positive and negative tweet data based on emoticon and utilize for model training. With this method, it may have noise in data, but it is cost effective.
2) **Sentimental path**; We enable a user can obtain insight at a glance by analyzing trajectory and sentiment together about the particular keyword. Usually trajectory and sentiment analysis are not used together, but we analyze and show a sentimental path.
3) **Deep learning on sentimental path**; We use recent deep learning techniques on analysis and prediction to improve accuracy. We utilize CNN to learn the sentiment analysis model and LSTM for the prediction model.

The remainder of this paper is organized as follows. In Section 2, we summarize related work. In Section 3, we explain our proposed system for analyzing and predicting users' sentimental trajectories for events. Section 4 shows the results of our validation experiments, and Section 5 concludes the paper.

2. Related Work

Social media sites enable users to create and share diverse contents anytime and anywhere by allowing users to freely create and share the contents. The free characteristics enable the sensing of the trends or changes in real life almost in real-time so that the trends and changes can be created, shared, and swiftly propagated as contents. As such, the use of social media sites has already penetrated into life to play the role of important contents in diverse fields. Accordingly, many studies to analyze and utilize social media contents are constantly conducted.

2.1 AsterixDB

One of the many important elements in the analysis of social media contents is how to efficiently manage and process massive contents. Since large amounts of contents are created and shared in real time, an efficient method of processing the contents is necessary. In addition, since the contents are not standardized like existing data, how to process the contents should be also considered.

For fast and efficient processing of social media contents, many sites and studies use MongoDB (Wei-ping, Ming-Xin, & Huan, 2011) that is based on NoSQL, which is freer than relational database. MongoDB is mainly used for managing social media contents, which is created in free formats because it has an advantage of efficiently managing dynamic documents such as JSON. Although MongoDB also can process data swiftly, it should process data queries faster to conduct social media contents analyses in real time. However, MongoDB has a disadvantage that it has relatively low responding speeds for JOIN queries. Since the speed for JOIN queries is also an important element to retrieve diverse pieces of information simultaneously and analyze the information, AsterixDB (Alsubaiee et al., 2014), which shows higher speed, is used.

AsterixDB is an open source BDMS (Big Data Management System) developed by Big Data group of UCI (University of California, Irvine) through an Apache project. Based on an object-oriented database and a NoSQL-style data mode that extended JSON, it has a characteristic of being flexible in processing semi-structured data. It also provides analyses of and queries for unstructured data. Clusters are configured easily for processing of massive data so that distributed processing is possible and it can be extended to at least 1,000 cores and at least 500 disks. In addition to the feed function to accommodate data created in real time, it supports diverse types of indexing such as B + tree, R tree, and inverted keywords (Alsubaiee et al., 2014).

2.2 Sentiment Analysis and Prediction

In social media sites, which are directly used in real life and enable the share of numerous pieces of information

in real time, contents created by users become major information. Contents that contain information or sentiments with photographs, short writings, and even only one word such as a hash tag are shared. The characteristic that information can be easily shared with more people with simple creations unlike existing blogs played a great role in enhancing users' use rates and the enhanced use rate and the characteristic that the contents are closely related to real life led to many studies on social media contents. Since social media contents are not standardized unlike existing traditional contents, for analysis, many pieces of information should be extracted from short writings (Senaratne et al., 2014; Tang et al., 2013).

The subject of sentiment analysis is usually made up of text. Therefore, NLP(Natural Language Processing) skills are necessary. There are dictionary based methods and machine learning based methods in sentiment analysis techniques. Sentiment analysis is organized in detail using two techniques (Medhat, Hassan, & Korashy, 2014).

SO-CAL(Semantic Orientation CALculator) which is sentiment analysis method using dictionary is proposed (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). To classify the sentiment of positive and negative, they used dictionary, which basically had numerical values for polarity for each word. They also tried to classify sentiment, taking into consideration any representation of negative or emphasis.

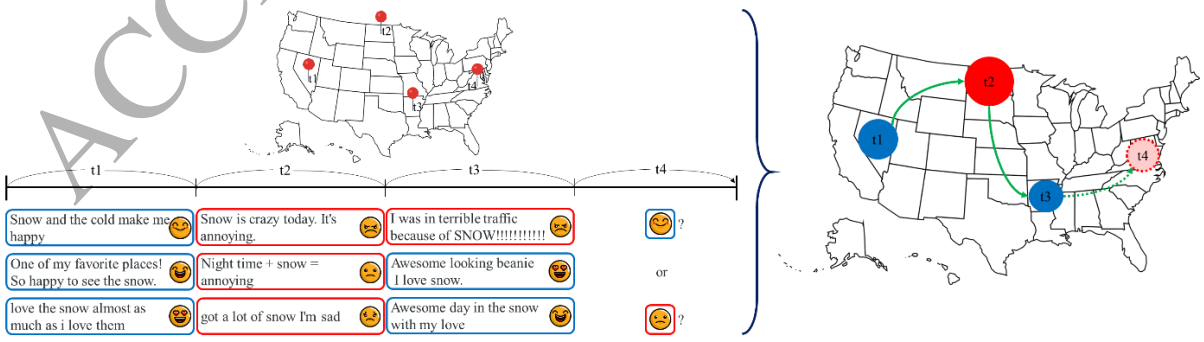
Machine learning based sentiment analysis gives a vectorized value to a word using a statistical method using word embedding, and trains a digitized sentence using machine learning or deep learning model. Traditional machine learning methods include SVM, Random Forest, and Naïve-Bayes, and deep learning models include CNN(Convolutional Neural Network), RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory), and GRU(Gated Recurrent Unit) (Nasrabadi, 2007; Pedregosa et al., 2011).

In traditional sentiment analysis, it is most important that the quality of the predefined dictionary or the labeled data affects on the analyzed results to distinguish between positive and negative text. First of all, we need to complement the process as it involved a process of labeling oneself. Therefore, we utilize the method that classify positive/negative based on emoticons (Do & Choi, 2015). In addition to reducing the cost that a person has to label oneself, it has also enhanced the accuracy of sentiment analysis by applying the recent deep learning techniques.

3. Proposed System

Social media contents have a characteristic of spreading rapidly because users can freely create and share them within the site. Users create numerous pieces of information or their opinions on information and share it with other users within the site. Through this process, social media contents respond sensitively to ever-changing social issues and natural disasters and the contents containing diverse information and sentiments are shared thereby rapidly spreading.

Sensitivity to social issues and high propagation speeds made the analysis of social media contents become important contents in grasping social issues. In this study, events in social media contents having such characteristics are identified and all of the locations, times, and texts of contents related to the relevant events are analyzed to analyze and visually show users' sentimental paths. In addition, a system that enables predicting users' sentiments based on the results of analysis is proposed.



<Figure 1> Motivating example of the snow

Figure 1 shows a motiving example of the snow for our proposed system. When a certain event has occurred, social media users create many contents containing information on the relevant event. Using this fact, sentiments on events remarkably frequently mentioned by users in a certain time zone (t) in a certain area

analyzed and the results of analysis are expressed on the map as sentimental paths. The sentiments in social media contents by time and place in the left part of Figure 1 are analyzed. Then, positive sentiments are indicated in blue and negative sentiments are indicated in red as shown in the right part of Figure 1. In addition, the sizes of nodes are made differently according to the intensity of sentiments so that the sentiments can be observed intuitively. The results of analysis are used as a model for prediction. As with the right part of Figure 1, the results of prediction can be shown using dotted lines based on the sentiment prediction model.

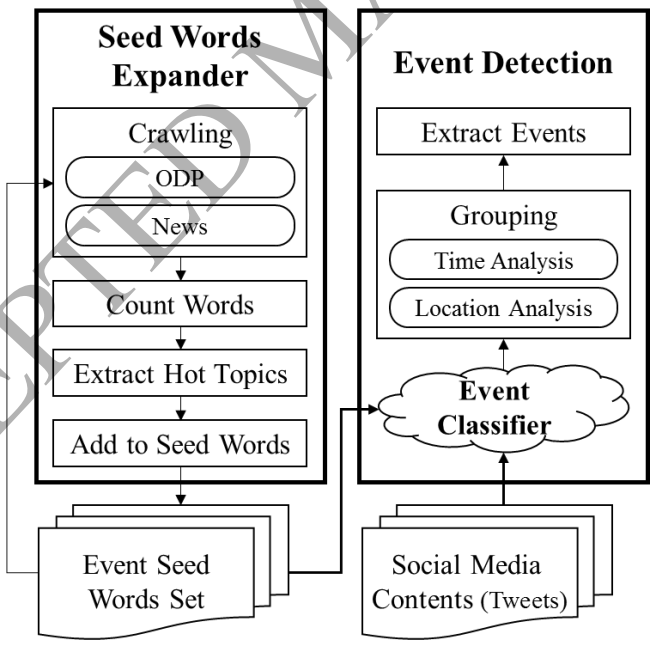
3.1 Event Detection and Trajectory Analysis

The social media contents that are continuously generated in real time show social issues and the daily lives of individuals. In particular, since contents regarding events that occurred in real life are shared and propagated rapidly, the results of analysis of what events occurred and how the events were propagated can be utilized for prediction together with the results of analysis of the events per se. We analyze social media contents to identify events and analyze and show the trajectories, which are propagation paths, of the relevant events.

3.1.1 Event Detection

Events can be largely divided into social events such as concerts and parties and disaster events such as typhoons and earthquakes. If events that occur over time and according to locations can be automatically identified, the results can be diversely utilized for disaster notices, event introductions, etc. A method that is the most frequently used to detect events is using sensors for event detection.

Since users firsthand share information on their daily lives, diverse events are shared in real time on social media sites. Since numerous contents poured on the timeline in real time have a characteristic of particularly intensively occurring at the time and place where a certain event has occurred, these contents can be used as direct event sensors. The occurrence of the event can be grasped by analyzing the time, location, and contents of the contents created.



<Figure 2> Process of event detection

Figure 2 shows the event detection process of the system proposed by us. Since the event seed words used for event detection must be continuously expanded, there is a separate seed words expander. The seed words collected and expanded through the seed words expander are used to judge whether or not there is information about events in social media contents. The contents where seed words exist are classify first and the time and place of occurrence of the relevant contents are analyzed to use the results for event detection.

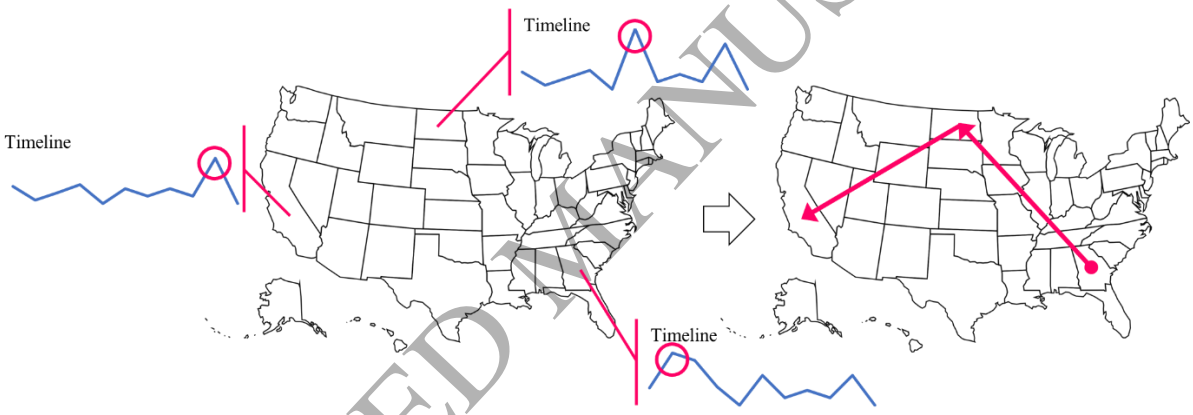
We use social media contents as a sensor for event detection. We detect events using the text containing the

contents of the contents that can be extracted from the contents firsthand created by users and the time and place of creation and utilize the results for trajectory analysis and sentiment analysis. To detect events based on texts, which keywords should be identified and extracted as keywords for the events is important. To identify events that contain events from among contents created in real time, those contents that contain certain keywords should be first classified. The keywords used in this case become the seed words for event detection.

The seed words used first for event detection are three words, {crime, disaster, accident} used in (Li, Lei, & Khadiwala, 2012). Seed words contain the keywords associated with the event. To collect these keyword, we collected keywords based on the ODP classification, crawled the articles on 'USAToday.com', which has the largest number of digital subscribers as of 2015, to expand seed words, and expanded the seed words using hot topic extraction.

3.1.2 Trajectory Analysis

Event trajectories are the results of analysis of the paths through which events are propagated through the analysis of the time and area where certain events have occurred. In fact, when a certain event or incident occurs, many related contents are created. In particular, in the case of Twitter, when a social issue or accident has occurred, related contents containing the relevant keyword are created quickly and propagated rapidly through retweet (RT) or sharing. By analyzing the time and location of the occurrence of social media contents containing specific keywords regarding the event, the paths through which the event is created and expanded can be identified through event path extraction (Song, Yoo, & Jeong, 2016; Yoo, Park, Song, & Jeong, 2017).



<Figure 3> An example of how to analyze the path

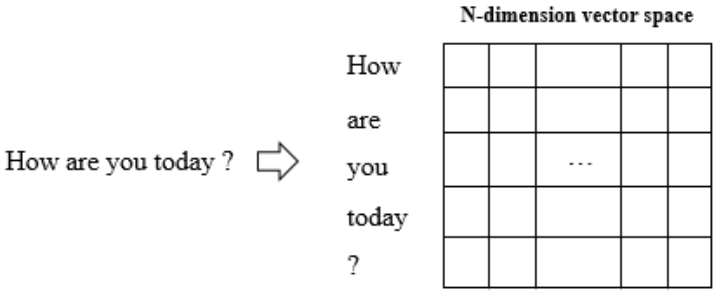
The trajectory analysis is conducted through a process of preprocessing of the collected social media contents followed by a process through which the paths are analyzed. Since the paths are analyzed based on the time and area of occurrence of contents, all contents should be grouped by area and the numbers of contents should be counted based on the time zones of occurrence of the contents grouped by area. Figure 3 shows an example of how to analyze the path. All the contents containing the keywords for the event are classified according to the areas where they occurred. The contents classified by area are classified again according to the time at which the contents occurred. When the contents have been arrayed in order of frequencies of occurrence in each time zone, the paths for keywords are extracted. The extracted paths used for sentimental trajectory analysis and prediction.

3.2 Sentiment Analysis and Prediction

Social media contents contain not only the information about the event that occurred but also the sentiments of the users who created them. The sentiments of users on a certain event are extracted through text analysis and the next sentimental path is predicted based on the results of extraction.

3.2.1 Sentiment Analysis

We use machine learning to analyze the sentiment of users for the events that occurred. Using TensorFlow, which is open source software, we undergo training for sentiment analysis. Users' sentiments can be immediately extracted by inputting texts out of users' contents through training.



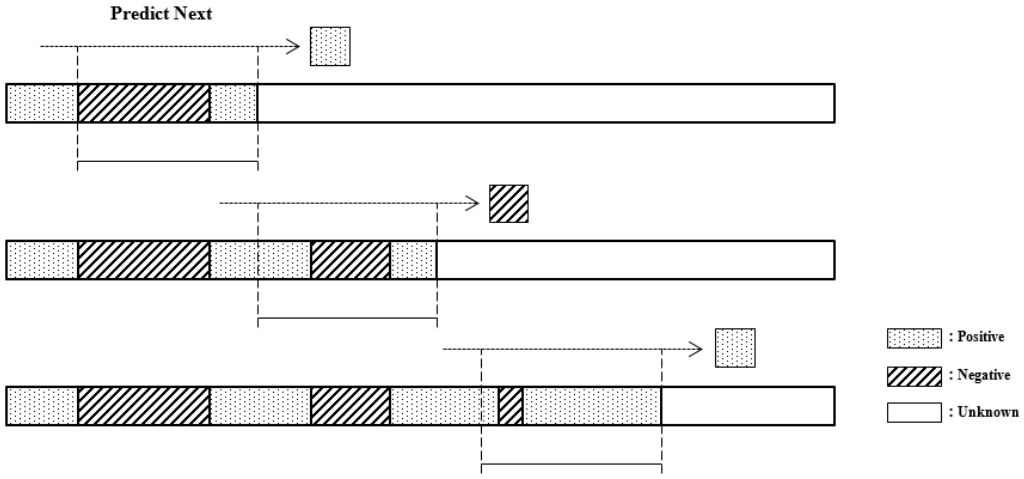
<Figure 4> Word vectorization

The sentiment classification model is trained using the Convolutional Neural Networks for Sentence Classification (CNN) which is a method proposed by (Yoon Kim, 2014) among diverse machine learning methods. CNN is a deep neural network model that shows very good performance in image classification problems. When this model is applied to sentiment classification, sentences are inputted as inputs of CNN like Figure 4. In this case, since all the words in sentences are vectorized using the word2vec algorithm, the sentences can be regarded as two-dimensional vectors like images. When we train our sentiment classification model, we used the dataset from 'sentiment 140 ("Sentiment140," 2017)' which contain 800,000 positive and 800,000 negative labeled data for training our model.

3.2.2 Sentiment Prediction

The results of analysis through our trained model can be utilized for prediction. How users' sentiments progress for events extracted in real time will be analyzed and the results of analysis will be utilized as a model for prediction of the next sentiment. By analyzing the sentimental path, the place where a certain event is expected to occur and the sentiment are predicted and shown in advance. The sentimental path prediction model enables diverse applications such as disaster prediction systems.

Sentimental paths are predicted by analyzing the sentiment of the contents for keywords for a certain event and applying the prediction algorithm to the results of analysis to predict the results of the next sentiment. Although sentiments can be highly accurately predicted when the machine learning algorithm is used, in cases where data on sentimental paths are not sufficient, the accuracy of the prediction model rather becomes much lower. Because of this problem, we predict sentimental paths through a calculation method using the weighted values rather than the machine learning algorithm.

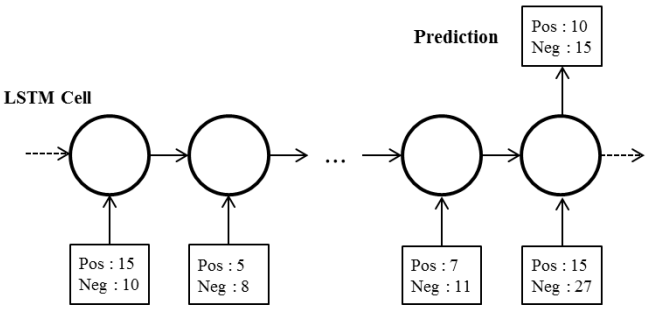


<Figure 5> The concept of time window

As Figure 5 shows, the concept of time window is used for sentimental path prediction. The size of the time window can be set in 7 day units like 7 days, 14 days, 21 days, and 28 days. We set the window size to 7 days for the experiment. When the time window has been observed as much as the set size, the next sentiment is predicted according to the resultant sentiment for the relevant period.

$$(1/window\_size)\sum sentiment_{pos} \times cont\_weight \qquad (1)$$

The equation 1 is used for sentimental path prediction. Sentiments as much as the set time window are analyzed to extract the means of positive and negative sentiments. In this case, the moving averages can be obtained by applying weights, and the resultant values used for the prediction of the next sentiments. The weights (*cont_weight*) are values between 0 and 1 for taking into account the continuity of the sentiments.



<Figure 6> LSTM for sentiment prediction

However, there is a limitation of using moving average to predict next sentiment. Because the average just makes counts in window smooth, it cannot be responsive to the dynamic movement of values. Therefore, we tried prediction using LSTM (Long Short-Term Memory) (Greff, Srivastava, Koutnik, Steunebrink, & Schmidhuber, 2016). Figure 6 shows LSTM for sentiment prediction. LSTM is a recurrent neural network architecture. It is well-suited to predict long time series data. Also, it has an advantage over traditional RNNs because of relative insensitivity about gap length (Greff et al., 2016; Hochreiter & Schmidhuber, 1997). So we conducted experiments, and used LSTM in our prediction since the experiment result showed that the method using LSTM was better than using moving average (Section 4.3.2)

4. Implementation and Experiments

In this study, we propose a system that can detect events sensed in social media contents, analyze and predict the sentimental paths of contents related to the relevant events. We will firsthand implement the proposed system to verify that we have done.

4.1 Implementation environment and data set

Ubuntu 14.04 (“Ubuntu 14.04,” 2017) was used to implement a system that detects events in social media contents and analyzes the propagation trajectories and sentiments. We used two Intel Xeon CPU E5-2620 v3 as CPUs and conducted implementation and experiments with specifications as follows; 500GB SSD, 32G memory, and GTX970 GPU. In particular, to process massive social media contents in real time AsterixDB was used as a database management system. Nvidia-docker, Tensorflow (“TensorFlow,” 2017), Python Flask (“Python Flask,” 2017), Nginx (“Nginx,” 2017), and Ruby on Rails (“Ruby on Rails,” 2017) were used for implementation.

In this study, data were collected from Twitter, which can be said to be a representative social media site, were used for implementation and experiments. Using streaming API provided by Twitter, we collected US tweet data for one month from April 1 to 30, 2016 and among approximately 40 million data, 10 million data extracted through outlier removal were used. We also utilized the data set provided in ‘sentiment 140’ to model the sentiment analysis. We used 800,000 positive and 800,000 negative data in modeling.

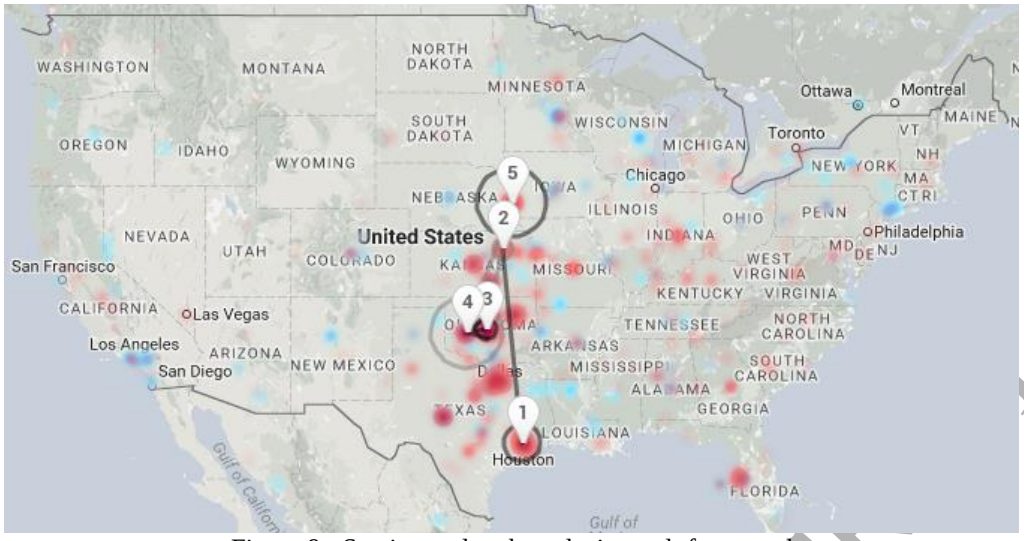
4.2 Implementation

We implemented a system that analyzes social media contents in real time to find and express events and analyzes and shows the trajectories and sentiments of the relevant events. We used the streaming API provided by Twitter to collect social media contents and used AsterixDB to manage collected contents efficiently. We also utilized the animation effects provided by Google Maps to visualize the analyzed results.

<Figure 7> Event detection result

Figure 7 shows the results of detection of events in real time. Based on the tweet data collected through the real-time streaming API, we analyzed the tweets that contain seed words using the proposed method. Accident related tweets identified in real time could be checked.

<Figure 8> Sentimental path analysis result for snow



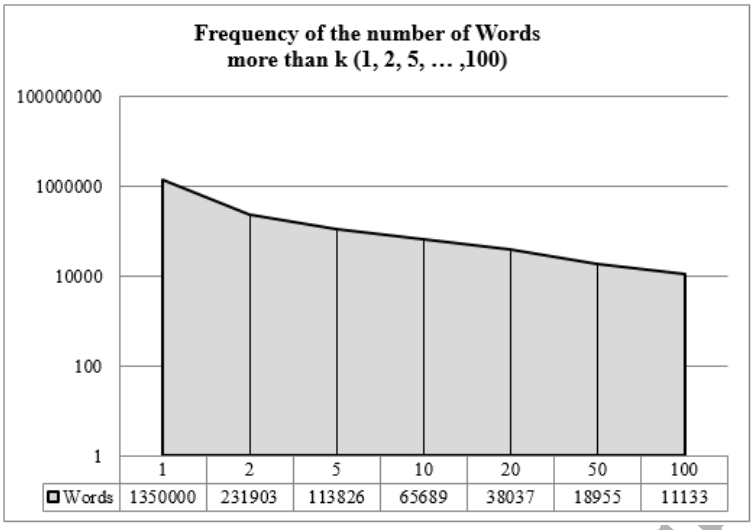
<Figure 9> Sentimental path analysis result for tornado

Figure 8 and Figure 9 show the sentimental path analysis results for snow and tornado, respectively. We used the flags to indicate the order of the trajectories. In addition, the intensity in each region was indicated by the size of the nodes by analyzing how many times the relevant keywords were mentioned in certain area. In fact, in the system, animations are drawn on the map according to the order. The data in the figures are tweet data from April 1 to 30, 2016. In fact, it snowed early April starting from the northeastern area to the northern area of the United States. In addition, a tornado occurred on April 3 in the Dallas area. As shown in the figure, the results of the trajectory analyzed in the proposed system reflect the actual movement paths.

In addition, the results of analysis of users' sentiments were expressed in blue and red using the hit maps. The red indicates negative sentiments while the blue indicates positive sentiments. The darker the color is, the greater the proportions of the relevant sentiments are. In the case of ‘snow’, relatively more negative sentiments occurred in area with heavy snow while more positive sentiments were shown in other areas with a positive sentiment ratio of 50.3% and a negative sentiment ratio of 49.7%. On the contrary, in the case of ‘tornado’, negative sentiments appeared more strongly with a negative sentiment ratio of 81.8%. Since the sentiments and paths for events occurred are shown together as such, the results of analysis can be visually identified much more easily.

4.3 Experiment

Diverse experiments were conducted to verify the accuracy of the proposed system. In particular, in the case of sentiment analysis, the accuracy of the sentiment analysis model should be verified because this model was trained. Except for the training data used in the sentiment model, we used 100,000 data for positive data and 100,000 data for negative data, a total of 200,000 data. Experiments on diverse variables such as the number of words in the vocabulary used in the sentiment model were conducted in order to find the optimum variables to our system.



<Figure 10> Word distribution

The distribution of unique words in our training dataset is as shown in the Figure 10. Y-axis is the frequency of the words that are shown more than number X-axis (1, 2, 5, …, 100) in Twitter data.

4.3.1 Experiment on Sentiment Analysis

We use distant supervision and CNN techniques to build sentiment analysis model. Whereas previously people had made model after labeling one by one, we collect positive/negative tweet data based on emoticons and use them in model training. Experiments have been carried out with variants of different variables to improve the accuracy of the sentiment analysis model. We have also validated the proposed method compared to traditional machine learning techniques.

Table 1. Accuracy according to vocabulary size

Vocabulary Size	Epoch 1	Epoch 2	Epoch 3	Epoch 4	Epoch 5
20000	0.8050	**0.8318**	0.7611	0.7802	0.7489
30000	0.7818	**0.8311**	0.7921	0.7695	0.7667
40000	0.8173	**0.8456**	0.8012	0.7937	0.7875
50000	0.8095	**0.8329**	0.7808	0.7746	0.7699
60000	0.7886	**0.8329**	0.7749	0.7842	0.7743

Table 1 shows the accuracy variation according to the word size of the sentiment model. The accuracy was calculated using the accuracy calculation method commonly used in information retrieval. It shows the accuracy of the word size from 1 to 5, respectively. As can be seen from the experiment results, the best result is shown when experiments were conducted two times regardless of the word size. From the third time and thereafter, overfitting has occurred so that the accuracy was rather lowered. Experiments were conducted while changing the variable diversely and according to the results, the optimum results were shown when the word size was 40,000, the number of filters was 128, and the batch size was 64. Therefore, the proposed system uses the sentiment analysis model applied with the aforementioned variables.

The proposed sentiment analysis is based on sentiment models. We use a sentiment analysis model that shows an accuracy of about 84% as with the abovementioned experimental results. Although training using more data is necessary to improve the accuracy, improved sentence analysis results can be obtained by considering the features of social media contents such as social relations.

We have experimented against traditional machine learning techniques to validate the proposed model. We used Naïve-Bayes, SVM, and Random Forest for the traditional machine learning models. We used the same datasets as the proposed model, and trained using the modules of Scikit-learn.

Table 2. Comparison result of sentiment analysis model with other machine learning methods

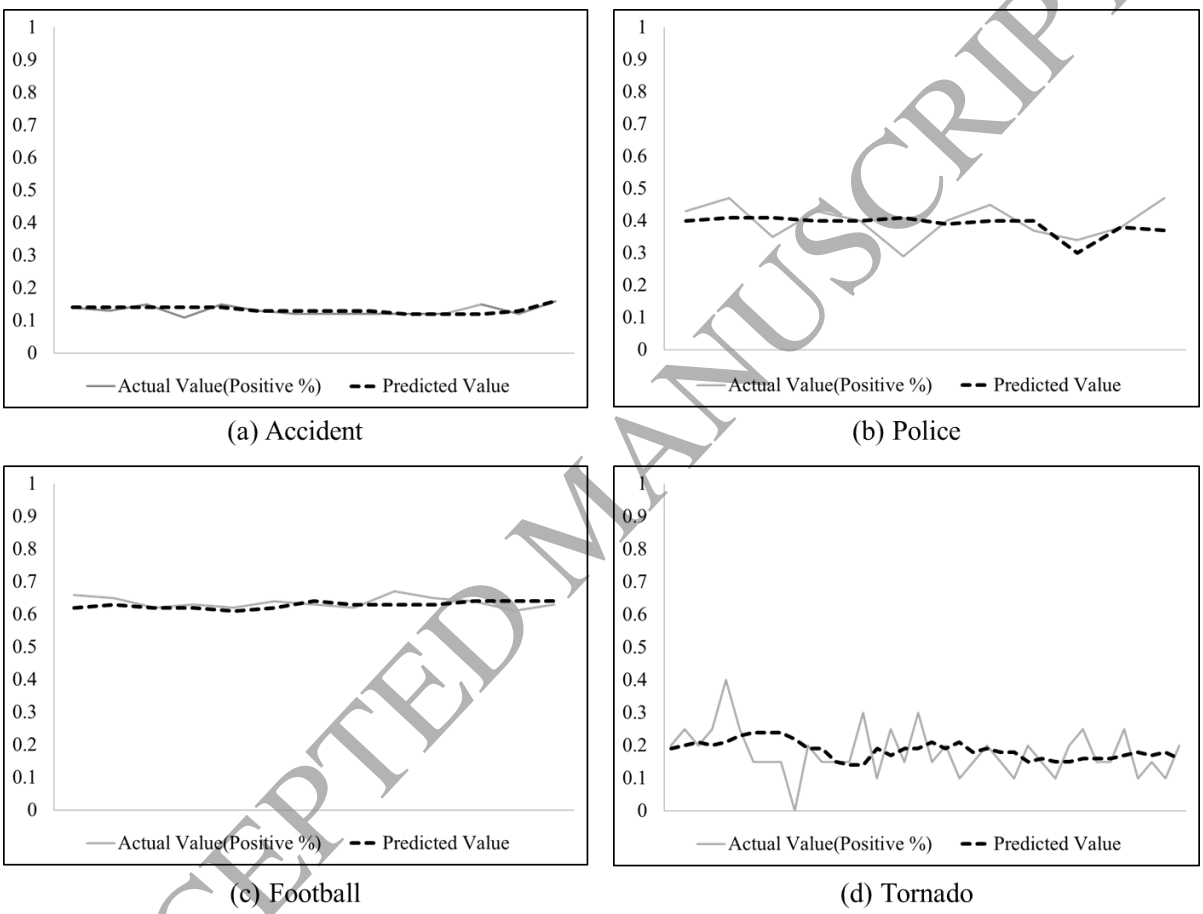
Model	Naïve-Bayes	SVM	Random Forest	Proposed Model

Precision	0.76	0.77	0.76	**0.839**
Recall	0.76	0.77	0.76	**0.845**
F-1 Score	0.75	0.77	0.76	**0.841**

Table 2 shows the result of comparison with the proposed model and the traditional machine learning methods. Although the traditional techniques show good performance, the best accuracy was found in the proposed sentiment analysis model using CNN.

4.3.2 Experiment for Sentiment Prediction

We propose a model to predict people's sentimental changes on the specific keyword. We propose moving average based method and LSTM based method. For each method, experiments were conducted to verify the sentiment prediction model.



<Figure 11> Comparison of actual resultant values for positive sentiments in sentimental path prediction

Figure 11 shows a graph of comparison of the percentage of actual positive sentiments obtained by analyzing the contents and the results of prediction. (a) shows search results for Accident, (b) shows search results for Police, (c) shows search results for Football, and (d) shows search results for Tornado. To analyze the accuracy of the sentiment prediction model, experiments were conducted with the keywords used as seed words for event detection. We set the window size to 7 days and used the US Twitter data for one month of April 2016 for the experiment. We analyzed whether the actual positive value matched the predicted result and showed the results in the graph of the figure. In the case of (d), although the accuracy is relatively low because only approximately 20 Tornado data were used, it can be seen that the accuracy can be higher if the quantity of data is sufficient through the accuracy of other keywords.

Table 3. Error rate according to window size

Window Size	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Accident	0.01	0.01	0.01	0.02	0.02	0.02	0.02

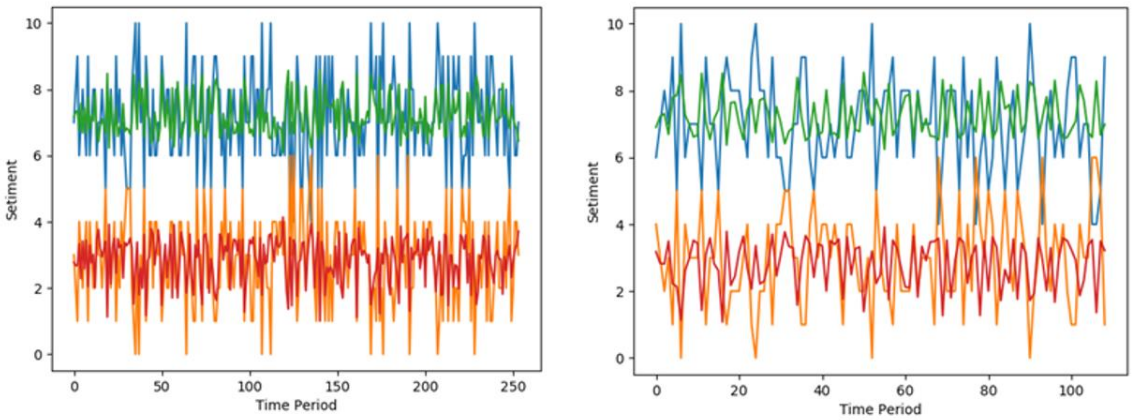
Police	0.05	0.05	0.05	0.05	0.04	0.04	0.04
Tornado	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Wreck	0.04	0.04	0.04	0.04	0.04	0.04	0.04
Snow	0.05	0.05	0.05	0.04	0.04	0.04	0.04
Flood	0.06	0.05	0.05	0.05	0.05	0.05	0.05
Death	0.02	0.02	0.02	0.02	0.02	0.01	0.01
Football	0.03	0.03	0.03	0.02	0.01	0.01	0.01
Baseball	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Concert	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Table 3 shows the error rates by keyword according to the window sizes. The error rates were shown to be relatively high for keywords with insufficient data. For more accurate comparison, in addition to the keywords used as seed words, we also conducted experiments with 'Football', 'Baseball', and 'Concert', which indicate social events, to show the usability of the sentiment prediction model.

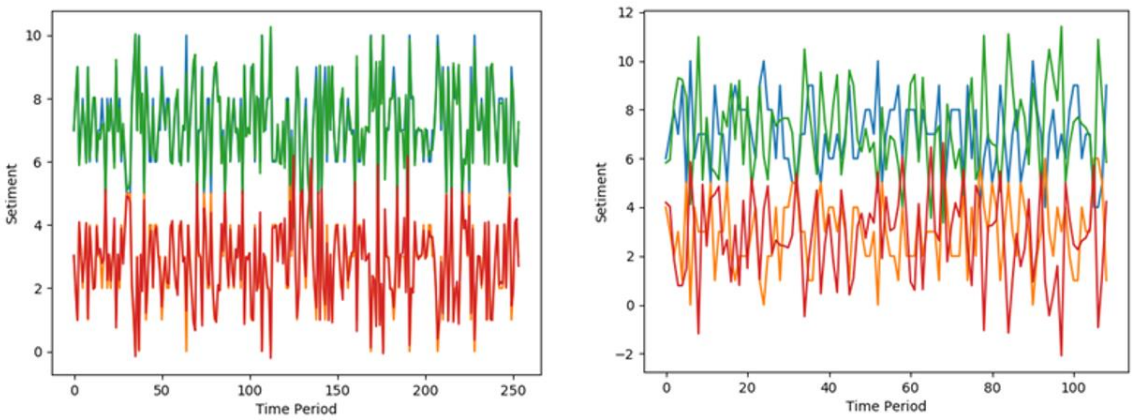
Table 4. RMSE of prediction results

Hidden dimension	**iterations**	**RMSE**
15	1500	1.6294
	2500	1.8933
	3500	2.2290
	4500	2.3379
25	1500	**1.5317**
	2500	1.9290
	3500	2.3371
	4500	2.4266
35	1500	1.6891
	2500	2.2465
	3500	2.4538
	4500	2.2784

Table 4 shows that increasing iteration at each hidden dimension sizes doesn't reduce the error. As we can see in the Figure 12, when we at the iteration 4500, our model can almost perfectly predict the training data, but RMSE (Root Mean Square Error) is increased.



(a) At iteration 1500



(b) At iteration 4500

<Figure 12> Training results; (a) at iteration 1500 (b) at iteration 4500. (Left side is using training data, and right side is using test data. Blue line is answer of positive, orange line is answer of negative, green line is prediction of positive, and red line is prediction of negative.)

Figure 12 shows that LSTM model can predict more dynamic pattern than the moving average method in the Figure 11. However, the accuracy of the model looks like not good because we just use one month twitter data to train our LSTM model. We expect that we can make a better model when we use lots of data.

5. Conclusion

We propose a system to analyze social media contents in real time while collecting and efficiently managing social media contents. We used AsterixDB to efficiently manage the social media contents in which data are accumulated in real time, and proposed and implemented a system that finds events in real time to analyze and predict users' sentimental paths.

We have collected and experimented social media contents for one month to test the proposed method. Because of the short collection period, there are limitations to the lack of training data for the sentiment analysis and prediction model. However, since multiple experiments shows potential of the proposed system, if we increase the amount of training data, and apply the trained model to real-time data, we can analyze and visualize the sentimental path on real-time.

Since the system can analyze shared social media contents freely used in real life in real time to find events related to disasters or accidents, and analyze and predict sentimental paths, the system can be utilized for disaster notice service for earthquakes and tsunamis or real time traffic accident informing service to immediately identify disasters and accident thereby reducing damage. In addition, since the system can be applied to social events too, it can be used for diverse marketing programs.

References

Alsubaiee, S., Altowim, Y., Altwaijry, H., & Behm, A. (2014). AsterixDB: A scalable, open source BDMS. *Proceedings of the VLDB Endowment*, *7*(14), 1905–1916. https://doi.org/10.14778/2733085.2733096

Cho, E., Myers, S., & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1082–1090. https://doi.org/10.1145/2020408.2020579

Do, H. J., & Choi, H. J. (2015). Korean Twitter Emotion Classification Using Automatically Built Emotion Lexicons and Fine-Grained Features. In *the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC 29)* (pp. 142–150). Deparment of Computer Science and Engineering Shanghai Jiao Tong University. Retrieved from http://hdl.handle.net/10203/204238

Feng, Z., & Zhu, Y. (2016). A Survey on Trajectory Data Mining: Techniques and Applications. *IEEE Access*, *4*, 2056–2067. https://doi.org/10.1109/ACCESS.2016.2553681

Greff, K., Srivastava, R. K., Koutnik, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–11. https://doi.org/10.1109/TNNLS.2016.2582924

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu, B., Jamali, M., & Ester, M. (2012). Learning the Strength of the Factors Influencing User Behavior in Online Social Networks. In *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 368–375). IEEE. https://doi.org/10.1109/ASONAM.2012.67

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. Retrieved from http://arxiv.org/abs/1408.5882

Kim, Y., & Moon, I. (2011). A study on algorithm for selection priority of contents in social network service. *Proceedings of the Korean Institute of Information and Commucation Sciences Conference*. Retrieved from http://society.kisti.re.kr/sv/SV_svpsbs03VR.do?method=detail&menuid=1&subid=11&cn2=HOJBAV_2011_y2011m10a_753

Lawton, G. (2001). Knowledge management: Ready for prime time? *Computer*, *34*(2), 12–14. Retrieved from http://dl.acm.org/citation.cfm?id=621644

Li, R., Lei, K., & Khadiwala, R. (2012). Tedas: A twitter-based event detection and analysis system. *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*. https://doi.org/10.1109/ICDE.2012.125

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/J.ASEJ.2014.04.011

Nasrabadi, N. M. (2007). Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, *16*(4), 49901. https://doi.org/10.1117/1.2819119

Nginx. (2017). Retrieved July 5, 2017, from https://www.nginx.com/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830. Retrieved from http://www.jmlr.org/papers/v12/pedregosa11a.html

Python Flask. (2017). Retrieved July 5, 2017, from http://flask.pocoo.org/

Ruby on Rails. (2017). Retrieved July 5, 2017, from http://rubyonrails.org/

Scellato, S., Noulas, A., Lambiotte, R., & Mascolo, C. (2011). Socio-spatial properties of online location-based social networks. *Proceedings of the Fifth International Conference on Weblogs and Social Media*. Retrieved from https://www.cl.cam.ac.uk/research/srg/netos/papers/icwsm2011_sociospatial.pdf

Senaratne, H., Bröring, A., & Schreck, T. (2014). Moving on Twitter: Using episodic hotspot and drift analysis to detect and characterise spatial trajectories. *Proceedings of the 7th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, 23–30. https://doi.org/10.1145/2755492.2755497

Sentiment140. (2017). Retrieved July 5, 2017, from http://www.sentiment140.com/

Song, J., Yoo, S., & Jeong, O. (2016). AsterixDB based Keyword Trajectory Analyzing System. In *2016 Korea Computer Conference* (pp. 317–319). Retrieved from http://insight.dbpia.co.kr/article/related.do?nodeId=NODE07017485

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Tang, J., Chang, Y., & Liu, H. (2013). Mining social media with social theories: a survey. *ACM SIGKDD Explorations Newsletter*, *15*(2), 20–29. https://doi.org/10.1145/2641190.2641195

TensorFlow. (2017). Retrieved July 5, 2017, from https://www.tensorflow.org/

Ubuntu 14.04. (2017). Retrieved July 5, 2017, from https://www.ubuntu.com/

Wei-ping, Z., Ming-Xin, L., & Huan, C. (2011). Using MongoDB to implement textbook management system instead of MySQL. *Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on*. https://doi.org/10.1109/ICCSN.2011.6013720

Yoo, S., Park, T., Song, J., & Jeong, O. (2017). A trajectory analysis system for social media contents using AsterixDB. In *Proceedings of the 11th International Conference on Ubiquitous Information Management and Communication - IMCOM '17* (pp. 1–6). New York, New York, USA: ACM Press. https://doi.org/10.1145/3022227.3022272

Zafarani, R., Abbasi, M., & Liu, H. (2014). *Social media mining: an introduction*. Cambridge University Press. Retrieved from https://books.google.co.kr/books?hl=ko&lr=&id=H9FkAwAAQBAJ&oi=fnd&pg=PR15&dq=Social+media+mining:+an+introduction&ots=cV3ervIVO8&sig=JhtBi0Y5jADm65TpwBlKGpDLQA4

Zheng, Y., & Zhou, X. (2011). *Computing with spatial trajectories*. Springer. Retrieved from https://books.google.co.kr/books?hl=ko&lr=&id=JShQJF23xBgC&oi=fnd&pg=PR3&dq=Computing+with+spatial+trajectories&ots=6MYghubicX&sig=SQUVmEuAGSO1ZLCNar6MZ6mOOvw