Twitter Sentiment Analysis using Machine Learning

Won Park, Youngin You, and Kyungho Lee*
Center for Information Security Technologies, Korea University, Seoul, Korea
{totoscawon, crenius, kevinlee}@korea.ac.kr

Abstract

As the number of social media users is being higher, many people are sharing various opinions and each country's real-time situation by online. Also, the influence of online information is increasing to such an extent that the individual's actual behavior or situation can be estimated. In this situation, researches to analyze through social media are being actively carried out in order to identify problems in real life. In this research, we proved that we can infer actual behavior or situation based on individual social media activities. This research focused on the Twitter platform that is actively used to express individual emotions in social media platforms. We analyzed tweets of Donald Trump and Hillary Clinton who were the 45th presidential candidates of the United States of America. Several methodologies like sentiment analysis, topic modeling, and machine learning were used to prove correlation between Donald Trump's tweets and his behavior. Through experiment, it proved not only we can adjust classification and clustering algorithms but also Decision Tree was the most accurate algorithm. Finally, we proposed the possibility of applying the above method to a system for detecting anomaly symptoms by concentrating on negative messages. It is expected to provide social media users with sufficient awareness of online activities.

Keywords: Donald Trump, Machine learning, Sentiment analysis, Topic modeling

1 Introduction

Donald Trump and Hillary Clinton, who were candidates for the 45th presidential election in the U.S., communicated with the world not only on the field election campaign but also through social media activities. In particular, Trump was a person who frequently used emotional and straightforward expressions during the presidential election and attracted the attention of the press and the public. His words and actions have not only made many issues, but also have been reported several times through the media in tough expressions and criticism of their opponents through Twitter. So we had a question as to whether Donald Trump's tweets reflect his behavior. The reason why we focused negative tweets was that his negative tweets show his speech and behavior well, and analyzing negative data is closely related to analyzing social problems. Recently, many organizations are collecting various information aiming at preventing negative behavior. There are many examples like closed circuit television (CCTV), collecting computer usage logs to prevent insider threat, disaster alert system collecting social media data, and so on. Of course, the way to prevent problems through these methods is not free from privacy issues, but it can't be denied that it contributes a lot to reduce damage by blocking the risk factors in advance. Likewise, this paper's main contribution is to propose the possibilities of implementing anomaly detection by collecting social media data. To prove this, Natural Language ToolKit (NLTK) scored the degree of sentiments about the tweets of the candidates and analyzed their association with actual behavior. In addition, topic modeling was used to extract topics in tweets to enable keyword-based analysis. Finally, we used Waikato Environment for Knowledge Analysis (WEKA) to show possibility of adapting machine learning algorithm.

Research Briefs on Information & Communication Technology Evolution (ReBICTE), Vol. 3, Article No. 9 (October 15, 2017)

*Corresponding author: Kyungho Lee, Center for Information Security Technologies, Korea University, Seoul, 02841, Korea, Tel: +82-2-3290-3937

2 Related Works

Sentiment analysis and topic modeling have been studied in the past to analyze human sentiments expressed on the web. Lu [9] used multi-aspect sentiment analysis and topic modeling as an example of restaurant evaluation. Nguyen and Shirai [11] designed a model that predicts stock price volatility using social media. Li [8] extracted the emotional keywords by applying the dependency sentiment LDA technique. Baek et al. [2] suggested ways to improve the problem of difficulty in processing high-level information about retrieving data using visual information. They proposed an object retrieval system based on KANSEI word dictionary. When a query is provided to the system, a shape matching the corresponding word is specified, and an image matches to the corresponding shape is retrieved from the DB and is provided. Although the above studies are effective in analyzing sentences, they are limited to analyzing correlation between emotion and behavior. Also, as the amount of information on the web grows excessively, the necessity of researching the summarization technique of text has emerged as a method for acquiring necessary information. As a research for this purpose, Lynn, Choi, and Kim [10] suggested a methodology for extraction summarization system. It was consisted by extracting keywords using TPDG model and constructing a lexical chain to generate summaries. And it was proved that the analysis of the acquired information could be performed with improved performance.

Especially, there were researches that proved the relationship between Twitter and real-world events. Hu, Farnham, and Talamadupula analyzed the average tweets per hour, hashtags, and retweets (RT) by two types of events (engaged or not engaged in an event). This proves that people tend to communicate more with others through social media activities when they are engaging events. In addition, they examined people were much more active in field of politics than technology, entertainment, and sports [6]. Meanwhile, a research showed that tweets reflects real-world's political aspect in example of German federal election. Approximately 100,000 tweets were collected to prove how the tweets affected election result [14]. In short, the topic modeling technique is applied to extract the keywords on tweets, and the sentimental score of each tweet is applied to understand the relationship between the sentiments of the collected tweets and the actual behavior.

3 Analysis Techniques

There are three methods that we used in this research: machine learning, NLTK, and topic modeling. All methods are correlated each other as a way to analyze Donald Trump's tweets. Also, there are several machine learning algorithms briefly in this section.

3.1 Machine Learning

Accurate analysis is possible when we remove unnecessary words and expressions included in tweets from the viewpoint of linguistic analysis. In this research, we used classification and clustering algorithms implemented in WEKA's. In classification algorithm, Naive Bayes is a kind of generative model to simplify learning process by minimizing the number of parameters that is necessary to learning [13]. Support Vector Machine (SVM) is to measure each group's distance of data and distinguish boundary, which is called 'Optimal Hyper Plane'. Joachims [7] introduced SVM as a way of text categorization. For Linear algorithm, Shalev-Shwartz and Ben-David [13] showed its predictors are good to analyze because it is intuitive, easy, and fit to natural problems. It is also used for the collected data by linear analysis (hyperplane, halfspace) to understand and predict the approximate trend. Decision Tree algorithm makes branches that can be classified from the root to make decisions about the learned data. Each 'leaf' on the 'tree' means a kind of status [12]. All of leaves are separated by branches that are find the best condition to satisfy. In clustering algorithm, Expectation Maximization (EM) is used to make cluster from incomplete data from iterative algorithm of E-step and M-step. Finally, SimpleKMeans is known for similar to EM, but it has a difference that makes k clusters and works to minimize variance of each cluster's distance.

3.2 Natural Language ToolKit (NLTK)

NLTK is a suite of program modules, data sets and tutorials supporting research and teaching in computational linguistics and natural language processing [3]. It works by dividing a given sentence into small units (corpus), learning and scoring a word-by-word evaluation index (vader). Sentiment is classified into three categories: positive, neutral, and negative. It depends on NLTK's compound score, which is in the range of -1 $\sim$ 1.

3.3 Topic Modeling

Topic modeling was presented as a way to manage effectively and classify the vast amount of information that flows through the web every day. Using this technique, it is possible to classify keywords and derive topics from unstructured documents [4]. Latent Dirichlet Allocation (LDA) is a good algorithm to reduce the dimensionality of the data and creates semantically consistent topics. LDA assumes that the words in the document are generated by combining the word distribution of the topic in the document with the topic distribution of the document.

4 Experimental Result

The purpose of the experiment is to prove that 'Donald Trump's tweets are related to his actual behavior. To prove this, emotion analysis, topic modeling, and machine learning methodology were performed. Also, to analyze the collected data numerically, the average score for each day was calculated, and the number of tweets by emotion level and the ratio of negative tweets among the daily tweets were calculated and plotted. The data thus obtained are classified according to the criteria established by themselves, and dates and keywords matching the criteria are selected. Analyzes the frequency of the keywords extracted from the tweets of the date, and verifies the correspondence by comparing with the actual events. In the process, we also conducted a comparison with Hillary Clinton to analyze the impact of Donald Trump's tweets. As a result, we found that Trump's negative tweet showed a lot of similarities with the behaviors seen in reality, and that it can be applied to machine learning.

4.1 Experimental setup

In this research, we crawled 4,000 of Donald Trump's tweets (@realdonaldtrump) created in 2016 (U.S. local time). Also, we used programs (crawler, NLTK, and topic modeling tool) that are made of open sourced Python code. After crawling, preprocessed unessential data like abbreviated link URL (e.g., https://t.co/Zt8TH-PtAn8) and some punctuation that can't compatible with WEKA (e.g., '@' in user mentioning expression (@username), double quotes (“”) and single quotes (‘’)), several parts of sentence (pronouns, prepositions, articles). Etc. After extracting keywords using topic modeling, we classified keywords into several categories to know which areas the candidates interested in. Next, we analyzed each candidate's online impact by comparing interaction with people (Like, RT, Mentions). And then, we made a dataset (3,500 tweets) and a testset (500 tweets) including tweet, compound, sentiment (positive, negative, neutral). The test set, which accounts for 10% of the total data, was consisted of randomly selected data from the dataset. Finally, converted dataset into .csv file to adapt WEKA and compare results of each algorithm of classification and clustering. We could get result values in classification (accuracy, precision, recall, and F-measure) and clustering (accuracy and ratio of each cluster). Figure 1 shows this research's entire architecture of experiment. It followed machine learning's typical flow and we inserted step of associating with real event.

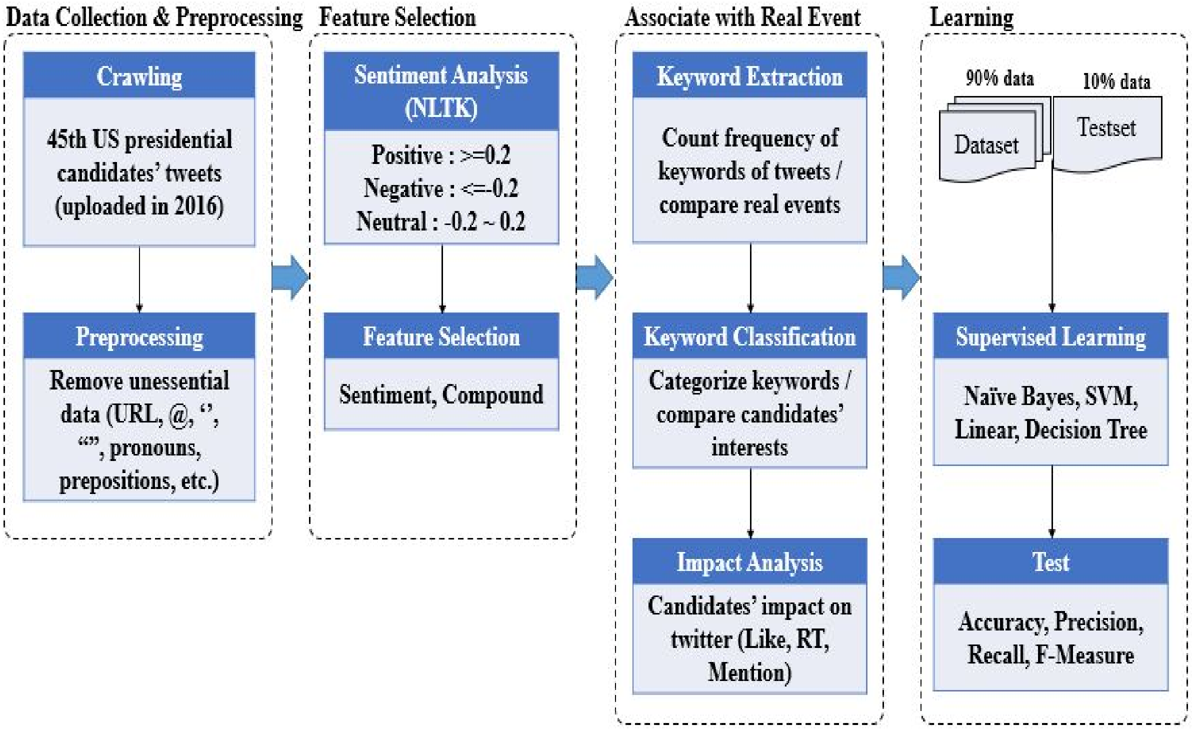
Data Collection & Preprocessing
Crawling
45th US presidential candidates' tweets (uploaded in 2016)
Preprocessing
Remove unessential data (URL, @, ', "", pronouns, prepositions, etc.)
Feature Selection
Sentiment Analysis (NLTK)
Positive : >=0.2
Negative : <=-0.2
Neutral : -0.2 ~ 0.2
Feature Selection
Sentiment, Compound
Associate with Real Event
Keyword Extraction
Count frequency of keywords of tweets / compare real events
Keyword Classification
Categorize keywords / compare candidates' interests
Impact Analysis
Candidates' impact on twitter (Like, RT, Mention)
Learning
90% data
10% data
Dataset
Testset
Supervised Learning
Naïve Bayes, SVM, Linear, Decision Tree
Test
Accuracy, Precision, Recall, F-Measure

Figure 1: Architecture of experiment

4.2 Daily sentiment analysis and keyword extraction

In order to compare the relationship between emotion and actual behavior, graphs were expressed on the basis of the daily average compound score and the number and ratio of positive / neutral / negative tweets. The sentimental criteria of [5] was converted and applied to this research. That is, [5] designated the sentiment score range from -5 to 5. It's positive when the score was higher than 1, negative when the score was lower than -1, and neutral otherwise. Likewise, it's positive when compound is 0.2 or more, negative when compound is less than -0.2, and neutral otherwise. The overall score was not significantly different (Trump: positive 54%, negative 29% / Clinton: positive 50%, negative 24%). However, when we consider the ratio of High-positive (compound score of over 0.6) and High-negative (less than -0.6), it proved that Trump had a stronger tendency in expression. Figure 2 and 3 shows the result of daily sentiment analysis of Trump's tweets from May to August. Each graph has several points that satisfy the conditions defined in this paper. We explained two cases of points at 4.2.1, 4.2.2. First case is about conflict between Trump and New York Times related to a woman affair. Second case is about blaming Hillary Clinton's economic issues. We searched news on Google regardless of media. We have specified a one-week period as a search range, including certain points, because we assumed that social media activity and actual action are able to be occurred immediately and also occurred at intervals of time.

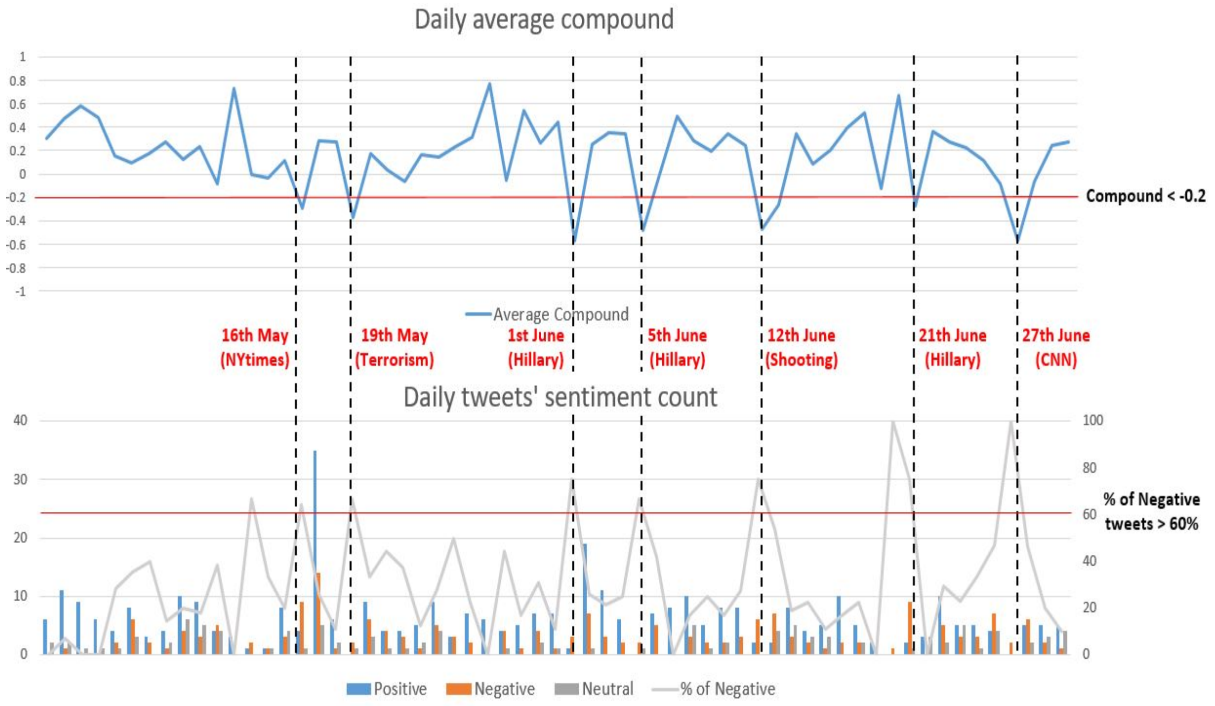

Figure 2: Daily sentiment analysis of Trump's tweets in May and June 2016

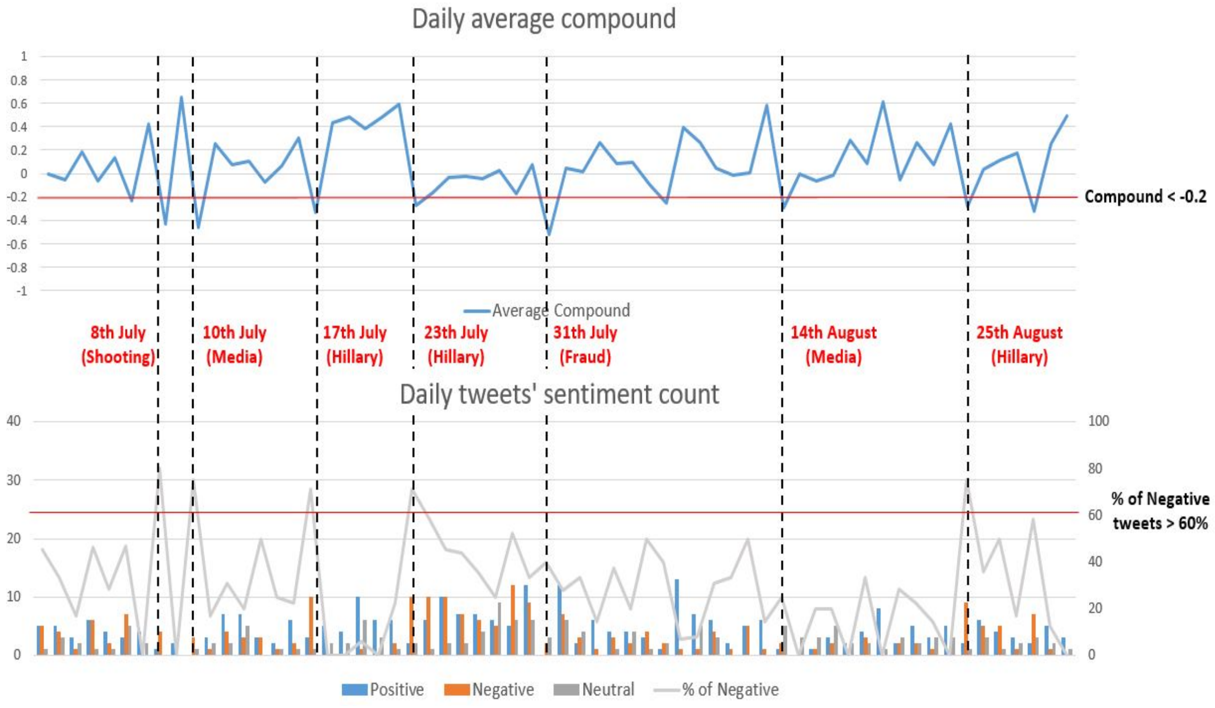

Figure 3: Daily sentiment analysis of Trump's tweets in July and August 2016

4.2.1 Keywords (16th May, 2016)

Compound is -0.2995 (Negative), and the rate of negative tweets is 64%. The most frequently mentioned words were ‘nytimes’, ‘story’, ‘failing’, and ‘woman’. These keywords are relevant with the conflict against New York Times. After being reported on Trump’s female affairs in the New York Times on May 14, Trump rebelled against it and wrote tweets to criticize they are ‘dishonest’. The following tweets also show that he made several conflicts with the media. ABC news reported Trump’s behavior that he was threatening to sue the New York Times and writing twitterstorm (2016.5.17.). The average compound of overall tweets that contain the keyword ‘nytimes’ is -0.3901, which suggests that Trump has a negative view of New York Times.

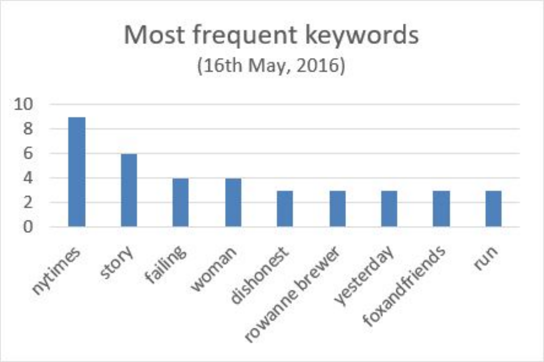

Figure 4: Most frequent keywords on 16th May, 2016

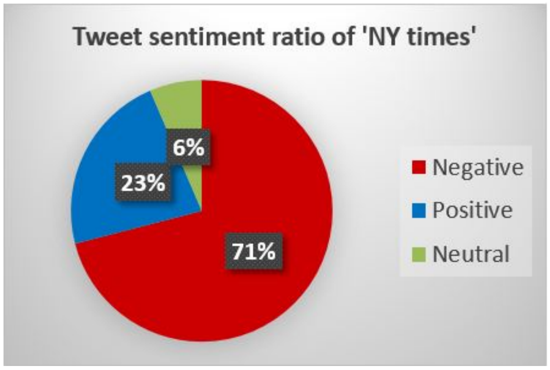

Figure 5: Sentiment ratio of keyword ‘NY-times’

4.2.2 Keywords (21th June, 2016)

Compound is -0.4735 (Negative), and the rate of negative tweets is 75%. The most frequently mentioned words were ‘hillary’, ‘great’, ‘crooked’, and ‘economy’. This shows that Trump and Clinton are criticizing with each other’s economic issues. Especially, Trump criticized Clinton at interview with CNN, mentioning ‘All of the money she’s raising, that’s blood money (2016.6.18.).’ The average compound of overall tweets that contain the keyword ‘hillary’ is -0.1402, which suggests that Trump has a relatively negative view of Hillary Clinton.

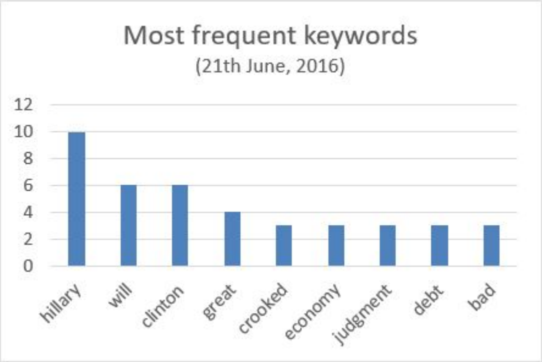

Figure 6: Most frequent keywords on 21th June, 2016

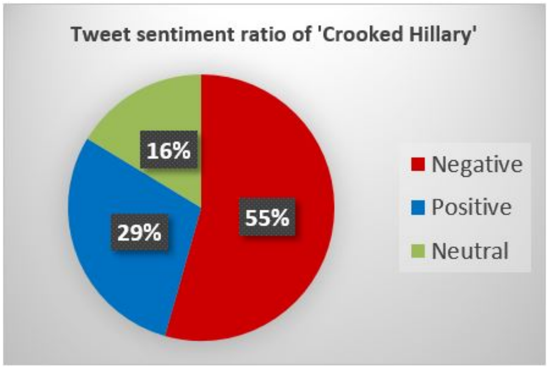

Figure 7: Sentiment ratio of keyword ‘Crooked Hillary’

4.3 Comparison of Keywords

To compare the interests of the two candidates, keywords were extracted from the collected tweets. There were similar interests because of the election in 2016. However, Trump was interested in media and employment, while Clinton gave priority to people. Media-related keywords were mentioned because Trump often conflicted with the media. And, because one of his commitments was employment expansion, keywords for employment were mentioned. On the other hand, Clinton frequently used keywords ‘Americans’, ‘families’, and ‘women’, as can be seen from the slogan ‘Stronger Together’.

Table 1: Frequently used Keywords

Trump			**Clinton**		
Category	Keywords	Count	Category	Keywords	Count
Candidates	Trump, RealDonaldTrump	1084	Candidates	Trump, Donald	1582
	Hillary, Clinton, crooked	1109		Hillary, Clinton, HillaryClinton	1319
Emotions	thank	644	Emotions	never	112
	great	525		good	101
	bad	124		believe	92
Commitments	MAGA, MakeGreatAmericaAgain	403	Commitments	plan	125
	AmericaFirst	96		going	113
	DrainTheSwamp	88		tax	106
	ImWithYou	86		together	102
Media	media	115	**People**	American(s)	234
	CNN	113		family(s)	192
	foxnews	92		women	165
Employment	jobs	105		help	101

4.4 Comparison of Candidates’ Impact on Twitter

We focused how many times of interaction (Like, RT, Mention) were made on each candidate’s Twitter. Because people who are engaged in events tend to be active in social media, they made more tweets, hashtags, and RTs [6]. From Table 2, the number of mentioning Trump (@realdonaldtrump) was much more than Clinton (@hillaryclinton)’s. Especially in Figure 8, the number mentioning Clinton was falling sharply after election day while Trump’s is similar to that of the election campaign. Likewise, the number of Like and RT to Trump’s tweets is more than Clinton’s. These statistics show interaction in social media is reflected to political aspect.

Table 2: Comparison No. of Likes and RTs

Candidate	No. of Tweets	Like		RT	
		Sum	Average	Sum	Average
Trump	4,000	78,386,034	19,596	33,421,976	8,355
Clinton	3,928	31,317,302	7,972	16,458,529	4,190

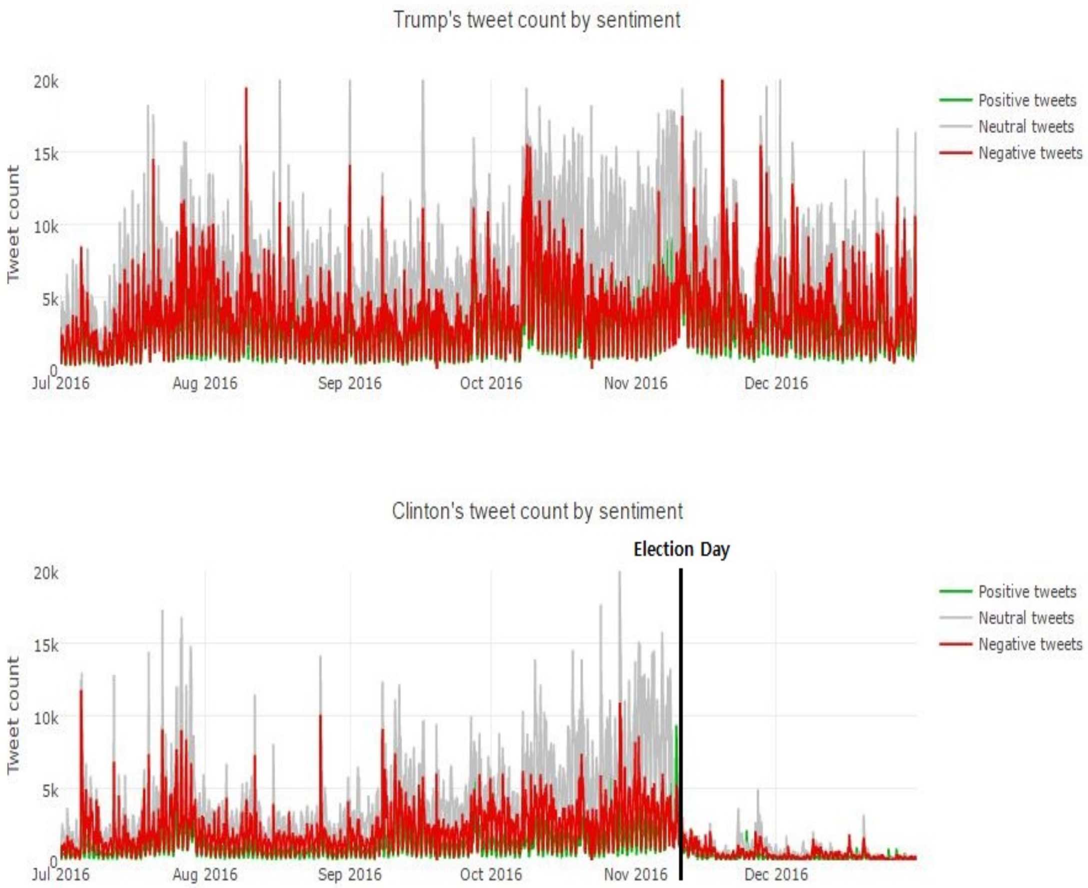

Figure 8: No. of Tweets Mentioned of Candidates [1]

4.5 Application of Machine Learning Algorithms

WEKA was used for tweet sentiment analysis using machine learning. Dataset and testset have elements as tweets, compound, and sentiment. The classification algorithms (Naive Bayes, SVM, Linear, and Decision Tree) and the clustering algorithms (EM, SimpleKMeans) were compared. Table 3 is an example of dataset for learning, Table 4 and Table 5 show the result of machine learning.

Table 3: Dataset example for WEKA learning

Tweet	Compound	Sentiment
Wow, Rowanne Brewer, the most prominently depicted woman in the failing nytimes story yesterday, was on foxandfriends saying Times lied	-0.2732	Negative
That was an amazing interview on foxandfriends - I hope the rest of the media picks it up to show how totally dishonest the nytimes is!	0.4587	Positive
Everyone is laughing at the nytimes for the lame hit piece they did on me and women. I gave them many names of women I helped-refused to use	0.1027	Neutral

Table 4: Result of classification by WEKA

Algorithm	Accuracy	Precision	Recall	F-Measure
Naive Bayes	97.4%	0.975	0.974	0.974
SVM	98.6%	0.986	0.986	0.986
Linear	96.6%	0.967	0.966	0.965
Decision Tree	100%	1	1	1

Table 5: Result of clustering by WEKA

Algorithm	Accuracy	Positive	Negative	Neutral
EM	95.725%	55%	28%	17%
SimpleKMeans	90.625%	45%	25%	31%

Each algorithm's results are slightly different and Decision Tree's result was the highest. Other algorithms show that the classification criterion is formed as a hyperplane, but the decision tree shows high accuracy because it can be classified as a kind of Yes or No question. Furthermore, classification algorithms showed relatively higher accuracy than the result of clustering algorithms. It shows that person's sentiment from social media can be adapted to classification and clustering in machine learning. But as a clustering, two algorithms showed different percentage in each sentiment.

5 Conclusion and Future Work

As the number of social media users increases, lots of information is being shared and disseminated online. In addition, minor information such as personal emotional state is also shared in a public form. However, it is possible that the actual behavior or situation of the individual can be leaked and the personal information issue can be caused. On the other hand, analyzed negative data from social media can help you detect abnormal signs, such as prevention of insider threats from organizations, disaster early warning and response. In this paper, it shows that there are information that can be leaked based on the personal information posted on Twitter and instills awareness about it. We also presented the possibility of data analysis to prevent problems. Through the above analysis, we could see that Trump's tweets were used as a way of expressing his feelings, and it's related his behavior that are exposed to press and people. Keywords that are frequently used showed correlation with the real events and sentiment can be analyzed using machine learning. This shows that the individual's social media activities can sufficiently simulate the person's actual behavior and situation. This research's results expected to provide social media users to have awareness of online behavior, as well as the possibility of applying an alert system. Every algorithm showed good result, but Decision Tree was the most accurate algorithm. However, we couldn't get satisfied result when we are in experiment about machine learning by each keyword. Therefore, further research will be extended to get sentiment of each keyword in tweets.

Acknowledgment

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the CPRC(Communications Policy Research Center) support program(IITP-2017-R1611402) supervised by the IITP(Institute for Information & communications Technology Promotion)

References

[1] Monkeylearn. `http://tarsier.monkeylearn.com/` [Online; Accessed on October 3, 2017].

[2] S. Baek, M. Hwang, M. Cho, C. Choi, and P. Kim. Object retrieval by query with sensibility based on the kansei-vocabulary scale. In *Proc. of the 2006 European Conference on Computer Vision Workshop on HCI, Graz, Austria, LNCS*, volume 3979, pages 109–119. Springer, Berlin, Heidelberg, May 2006.

[3] S. Bird. Nltk: the natural language toolkit. In *Proc. of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[4] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[5] S. Feng, J. S. Kang, P. Kuznetsova, and Y. Choi. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *Proc. of the 2013 ACL Conference, Sofia, Bulgaria*, pages 1774–1784, August 2013.

[6] Y. Hu, S. Farnham, and K. Talamadupula. Predicting user engagement on twitter with real-world events. In *Proc. of the 9th International AAAI Conference on Web and Social Media (ICWSM'15), Oxford, UK*, pages 168–178, May 2015.

[7] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proc. of the 10th European Conference on Machine learning (ECML'98), Chemnitz, Germany, LNCS*, volume 1398, pages 137–142. Springer, Berlin, Heidelberg, April 1998.

[8] F. Li, M. Huang, and X. Zhu. Sentiment analysis with global topics and local dependency. In *Proc. of the 24th AAAI Conference on Artificial Intelligence (AAAI'10), Atlanta, Georgia, USA*, pages 1371–1376, July 2010.

[9] B. Lu, M. Ott, C. Cardie, and B. K. Tsou. Multi-aspect sentiment analysis with topic models. In *Proc. of the 11th IEEE International Conference on Data Mining Workshops (ICDMW'11), Vancouver, Canada*, pages 81–88. IEEE, December 2011.

[10] H. M. Lynn, C. Choi, and P. Kim. An improved method of automatic text summarization for web contents using lexical chain with semantic-related terms. *Soft Computing*, pages 1–11, April 2007.

[11] T. H. Nguyen and K. Shirai. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proc. of the 2015 ACL Conference, Beijing, China*, pages 1354–1364, July 2015.

[12] S. Schrauwen. Machine learning approaches to sentiment analysis using the dutch netlog corpus. Technical report, Computational Linguistics and Psycholinguistics Research Center, July 2010.

[13] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[14] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *Proc. of the 4th International AAAI Conference on Web and Social Media (ICWSM'10), Washington, D.C., USA*, 10(1):178–185, May 2010.

Author Biography

Won Park received his B.S. degree in Computer Science from Sangmyung University in 2011 and is currently in his M.S. degree at the Graduate School of Information Security, Korea University. His research interests include machine learning, natural language processing, analyzing malicious code and incident response. He is also a member of Risk Management Laboratory in Korea University.

Youngin You received M.S. degree from Korea University. He is now a Ph.D Candidate at the Graduate School of Information Security, Korea University. Currently he is a member of Risk Management Laboratory in Korea University. His Research interests include Information Security Management System, Security Maturity and Privacy.

Kyungho Lee received his Ph.D degree from Korea University. He is now a professor in the Graduate School of Information Security, Korea University, and has been leading the Risk Management Laboratory in Korea University since 2012. He was a former CISO at NHN Corporation, and CEO of SecuBase Corporation.