Sentiment Analysis in Microblogs Using HMMs with Syntactic and Sentimental Information

Noo-Ri Kim, Kyoungmin Kim, and Jee-Hyong Lee

Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Korea

Ijfis

Abstract

In this paper, we propose an approach for sentiment analysis in microblogs that learns patterns of syntactic and sentimental word transitions. Because sentences are sequences of words, we can more accurately analyze sentiments by properly modeling the sequential patterns of words in sentimental sentences. However, most previous research has focused on just extending feature sets using n-grams, POS tags, polarity lexicons, etc., without considering sequential patterns. Our proposed approach first identifies groups of words that have similar syntactic and sentimental roles, called SIGs (similar syntactic and sentimental information groups). We then build HMMs using the SIGs as hidden states for the initialization. The SIGs function as the prior knowledge of formative elements of sentimental sentences for HMMs. By using the SIGs, HMMs can start with informative hidden states and more precisely model the transition patterns of words in sentimental sentences with robust probability estimation. For the performance evaluation, we compare the proposed approach with existing ones using HCR dataset. The result shows that the proposed approach outperforms the previous ones in various performance measures.

Keywords: Sentiment analysis, Hidden Markov models, Gaussian mixture models, Syntactic and sentimental information

1. Introduction

Microblogging services have become popular communication tools among Internet users, and social network services (SNS) have been growing at a rapid pace. Users write microblogs to express their own thoughts and share information about various topics through SNS platforms. Thus, microblogs can reflect diverse user interests on trending topics such as news, products, and social events. Many researchers have sought ways to mine microblogs for user opinions on trending issues [1–7].

Sentiment analysis is one of the basic utility functions needed by various applications for documents. The purpose of sentiment analysis is to identify the sentiment polarity of a document as positive, negative, or neutral. Many approaches based on machine learning techniques have been proposed following the lead of Pang and Lee [8], who used machine learning algorithms to build classifiers from documents with manually annotated sentiment polarity.

Previous approaches have focused mainly on conventional documents, such as movie

Received: Dec. 13, 2017
 Revised : Dec. 25, 2017
 Accepted: Dec. 25, 2017

Correspondence to: Jee-Hyong Lee
 (john@skku.edu)
 ©The Korean Institute of Intelligent Systems

©This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

reviews, blogs, and news. They used machine learning algorithms in which a document is represented as a bag-of-words. In these approaches, documents usually consist of at least paragraph-length pieces of text. Moreover, the texts are relatively well-formed and coherent. Such approaches have indeed proven to be quite successful [8–10].

On the other hand, sentiment analysis of microblogs is much harder because of that medium's different characteristics. For example, Twitter, one of the most popular microblogging services, differs from conventional documents. The length of tweets is limited to 140 characters, but the word diversity in tweets is the same as in conventional documents [6]. This causes a sparseness problem that makes it hard to develop informative feature sets. To alleviate this problem, most studies specializing in the microblog sentiment analysis have added more features, such as n-grams, part-of-speech (POS) tags, or other polarity lexicons [1–3]. However, those features may worsen the sparseness problem. For example, here are two positive sentences:

This movie is very good.

This movie is not boring, it is not ordinary:).

Correctly classifying both sentences as positive requires bigrams, such as “not boring” and “not ordinary” as well as unigrams such as “good” as features for positive sentiment. To guarantee an accurate sentiment analysis, a huge number of bigrams would have to be added. This could also worsen the sparseness problem because the frequency of bigrams is quite low in microblogs. Hence, we here consider another way to improve the performance of sentiment analysis in microblogs.

In this paper, we propose a new sentiment analysis approach for microblogs using word sequences. Because sentences in microblogs are sequences of words, we can more accurately analyze sentiments by properly modeling sequential patterns of words in sentimental sentences. To take word sequences into account, we adopt hidden Markov models (HMMs), which are well-known for sequential learning [11, 12]. However, a simple application of HMMs might not guarantee good performance. Because HMMs model hidden states and transitions between them in a blind probabilistic way, they cannot be guaranteed to properly model word transitions in natural language sentences. Also, a given transition from word a to word b in tweets might be very infrequently observed because a large number of words appear in very short tweets, and that might limit the HMMs from producing robust probability estimations.

Therefore, instead of simply modeling word transitions, we model transitions with syntactic and sentimental information using syntactic and sentimental transition patterns in sentimental sentences. For example, some positive sentences have the form

(pronoun or noun)(neutral verb)(positive adjective)

or

(pronoun or noun)(neutral verb)(negation)(negative adjective).

“This movie is very good” is an example of the first form, and “This movie is not boring, it is not ordinary :)” is an example of the second form. Transitions frequently observed in positive sentences differ from those observed in negative sentences. For example, transitions from (neutral verb) to (positive adjective) or from (negation) to (negative adjective) will be observed with higher probabilities in positive sentences. On the contrary, such transitions are rarely found in negative sentences. If HMMs can model such transitions, they can produce accurate sentiment analysis.

To model syntactic and sentimental word transition patterns, we first identify groups of words with similar syntactic and sentimental roles. In the above examples, (pronoun or noun), (negation), (neutral verb), (positive adjective), and (negative adjective) are such groups; we call these groups SIGs (similar syntactic and sentimental information groups). To group words together, we first develop a feature set of words that indicates their syntactic and sentimental roles. Then, we group those words into SIGs with the developed feature set using Gaussian mixture models (GMMs) [13]. In the initialization, the SIGs are used as the hidden states of the HMMs. Because the SIGs have syntactic and sentimental information about words, the HMMs can more properly model the transition patterns found in sentimental sentences with robust probability estimations. For the performance evaluation, we provide experimental demonstrations and various comparisons with existing works. As a result, we show that our proposed approach can analyze the sentiment of microblogs more accurately and robustly than existing approaches.

The rest of this paper is organized as follows: In the next section, we describe the related work on sentiment analysis. In Section 3, we propose a new approach for sentiment analysis in microblogs. We analyze and compare the performance of the proposed approach with other existing works in Section 4. In the last section, we provide a conclusion.

2. Related Work

Sentiment analysis in microblogs attempts to identify and analyze the sentiment polarity of short and informal texts. A large amount of work has been conducted in this field [1–3, 14–22]. Earlier studies mainly focused on lexical resources. They assumed that the semantic orientation of a document is an averaged sum of the semantic orientations of the words. Those studies relied on predefined lexical resources to compute sentiment polarity scores and used those scores to determine the polarity of documents [10, 23]. However, open lexical resources are not always sufficient to handle sentiment scoring problems in various domains. Also, people often use new expressions, such as “thx” (thanks), and various emoticons such as “:)” in microblogs. Predefined lexical resources cannot capture the sentiment information of such words.

In recent years, most studies have addressed the sentiment analysis problem as a text classification task [2–4, 6, 7, 15, 18, 19]. Thus, researchers have built classifiers using a machine learning algorithm on a dataset with features such as unigrams, bigrams, etc. Building good classifiers requires the development of informative features. Many studies combined n-grams with other features, such as sentimental lexicons given by open lexical resources.

Kouloumpis et al. [3] investigated the utility of various features such as n-grams, POS tags, microblog features, etc. and compared them to find effective features. Bravo-Marquez et al. [4] combined several existing lexical resources to develop meta-level features such as the opinion strength of lexicons, emoticons, etc. and applied them to machine learning algorithms, such as support vector machines (SVMs) and naïve Bayes (NB). Mohammad et al. [18] also analyzed the relationship between various features and sentiment analysis performance. Speriosu et al. [7] exploited Twitter follower graphs as additional information with users, tweets, lexical features, etc. as its nodes. They then used maximum entropy classifiers in combination with the graph. Liu et al. [16] trained a sentimental model based on labeled data and then used noisy labeled data for smoothing. Matsumoto et al. [17] proposed extended n-gram features instead of using simple n-grams. They mined frequent sub-patterns from sentences by considering the syntactic information of words. Hutto and Gilbert [15] proposed a rule-based model for sentiment analysis from social media text. They created rules from a sentiment lexicon made by experts.

Those studies found that extending feature sets using lexical resources, other extra information, or expert knowledge

improved sentiment analysis. However, improving the performance requires a huge number of features to be added. Simply extending feature sets can cause the sparseness problem in feature vectors because although many words appear in tweets, fewer than 15 words are included in an average tweet [16]. Therefore, as the number of features increases, the sparseness level also increases, and the performance is bound to worsen.

To alleviate the sparseness problem, Saif et al. [6, 19] used named entities to reduce the number of unigrams, extracted sentiment-topic features from the unigrams, and used them to determine the polarity of microblogs. However, their performance is greatly affected by the quality of the named entity extraction process, and that performance depends on domains.

As mentioned above, most approaches have tried to extend features by combining various resources. However, sentences are sequences of words with certain patterns that carry the sentiment information. Therefore, just extending features is not enough to handle the problem of sentiment analysis in microblogs. Thus, we consider sequential patterns to more effectively model the sentiments of sentences.

3. Proposed Approach

3.1 Overview

In this paper, we propose an approach to analyze the sentiment of microblogs by learning the patterns of syntactic and sentimental word transitions. As we mentioned in Section 1, we can find some sequential patterns in sentimental sentences. For example, positive sentences frequently contain transitions from (neutral verb) to (positive adjective) or from (negation) to (negative adjective). We thus focus on groups of words that have similar roles from the syntactic and sentimental points of view. For example, “good,” “happy” and “nice” are adjectives with positive polarities, so they can be grouped together into a SIG of (positive adjective). Using SIGs, we model transitions between words. Figure 1 shows the basic idea of the proposed approach. Given the sentence “I feel happy,” we interpret it using the transitions between SIGs, (pronoun) to (positive verb) and (positive verb) to (positive adjective) instead of directly interpreting the sentence as transitions from “I” to “feel” and “feel” to “happy.”

The overall procedure of the proposed approach contains three steps: 1) feature selection, 2) identification of SIGs, and 3) HMM training and sentiment analysis.

Figure 1. Basic description of the proposed approach.

3.2 Feature Selection

Our proposed approach is based on sequences of unigrams. To reduce the number of unigrams observed, we choose them by using term frequency of all words and each class. We select top- n words for the whole word, positive words, and negative words, and remove redundant words.

3.3 Identification of SIGs

To identify SIGs, we develop a feature set for grouping unigrams. The feature set needs to reflect the syntactic and sentimental roles of the unigrams.

3.3.1 Syntactic features

Features with syntactic information include POS tags and contextual valence shifter types. POS tags are linguistic word categories generally defined by words' syntactic or morphological behavior in grammar. However, some words in microblogs belong to new syntactic categories, such as user, URL, emoticon, abbreviation, and punctuation. So, we add those as POS tags. Next, we focus on contextual valence shifters, such as not, hardly, rather, etc. Those words change the degree of the expressed sentiments to weaken or strengthen the base valence of the following terms. They are very important to sentiment analysis. We use Negations [24].

3.3.2 Sentimental feature

We use one feature for sentiment information: positive, negative, or neutral polarity of unigrams.

We develop syntactic-sentimental features by combining 26 syntactic features and 3 sentimental features to group unigrams. For example, adjectives can have sentimental polarities, so we

subdivide those into 'positive adjectives,' 'negative adjectives,' and 'neutral adjectives.' For words that do not have sentiment polarities, such as auxiliary verbs, determiners, and conjunctions, we do not subdivide. Because Negations do not have sentimental information, we use a total of 76 syntactic-sentimental features to group unigrams into SIGs.

We use TweetNLP, an English POS tagger specialized in Twitter data, to obtain the POS tags. Next, we use the valence shifter list of Polanyi and Zaenen [24] to identify the contextual valence shifters.

To determine the sentiment information of words, we use SentiWordNet [23]. It provides information on the polarity of words. If a word w is included in the lexical resource, its sentiment is decided. If w is not included in the lexical resource, our proposed approach uses WordNet to find a synset of w . Once the synset is found using the relations of synonyms, we obtain the sentiment scores of the words in the synset from SentiWordNet. Then we use their average score, $sentiScore(w)$, for the sentiment score of w , as shown in Eq. (1).

$$sentiScore(w) = \frac{\sum_{v \in synset(w)} sentiScore_{SWN}(v)}{|synset(w)|}. \quad (1)$$

In the equation, $sentiScore_{SWN}(v)$ is the sentiment score of v in SentiWordNet. If $sentiScore(w)$ is larger than 0, the polarity is *Positive*. If it is less than 0, it is *Negative*; otherwise, it is *Neutral*. If w is not included in WordNet, $sentiScore(w)$ is defined as zero.

We describe unigrams using their syntactic-sentimental features. The feature values are binary. If a unigram has been used in a certain syntactic-sentimental role in the training data, the value of the corresponding feature is set to 1. If not, the feature value is zero. For example, the positive training set contains, "I love a beautiful watch." and "we are in love!" The unigram 'love' is used as a positive verb in the former instance and a positive noun in the latter instance. The values of (*positive verb*) and (*positive noun*) for 'love' become 1. If the only syntactic-sentimental features are (*positive verb*, *neutral verb*, *negative verb*, *positive noun*, *neutral noun*, *negative noun*), then the feature vector of 'love' is represented as (1, 0, 0, 1, 0, 0).

Using such syntactic-sentimental feature vectors, we group unigrams into SIGs, which are then used as the hidden states of HMMs. Figure 2 shows the procedure for identifying SIGs. Because words can be used differently from syntactic and sentimental points of view depending on the sentiments expressed, we build different SIG sets for each polarity. For the positive HMM, we identify SIGs with positive instances in the training

Figure 2. The procedure for identifying SIGs.

sets. We first build positive syntactic-sentimental feature vectors of unigrams with positive instances. Then we apply GMMs to group unigrams into SIGs. We build SIGs for the negative HMM in a similar way. For example, ‘good’ is usually used as a positive adjective in positive sentences, but it is frequently used as a neutral noun in negative sentences. Therefore, ‘good’ will mainly belong to a SIG including positive adjectives in the positive HMM, but it will belong to a SIG including neutral nouns in the negative HMM. We can thus properly group words considering their usages in sentimental sentences.

3.4 HMM Training and Sentiment Analysis

In this paper, we use HMMs to build a sentiment analysis model. We use SIGs as the hidden states of HMMs and set the initial emission probabilities that words belong to SIGs as shown in Eq. (2).

$$P(o_t = v_k \mid q_t = SIG_j) = P(v_k \mid SIG_j). \quad (2)$$

In the equation, o_t and q_t are the symbols and the state at t , respectively, v_k is the k -th unigram, and SIG_j is the j -th SIG.

The transition probabilities are randomly initialized. Then, we apply the general training algorithm for HMMs, the Baum-Welch algorithm, to adjust the transition probabilities between hidden states and the emission probabilities of words at hidden states.

Once the positive and negative HMM classifiers, $\lambda_{Positive}$ and $\lambda_{Negative}$, have been created, the sentiment analysis is conducted. Given a sequence O , the class of O is determined using the likelihood of each model. The polarity with the maximum likelihood will be chosen as the label for the sequence O . The

maximum likelihood decision rule can be stated as Eq. (3).

$$\hat{y} = \operatorname{argmax}_{\lambda_c \in \Lambda} P(O \mid \lambda_c), \quad (3)$$

where $\Lambda = \{Positive, Negative\}$.

Because SIGs are groups of words with similar syntactic and sentimental roles, they can be regarded as formative elements of sentimental sentences. By using SIGs, HMMs start with informative hidden states that correspond to formative units of sentimental sentences. Compared with a random initialization, these HMMs can more exactly model formative units and the transitions between them. Thus, our proposed model can easily find the transition patterns specific in each polarity and robustly estimate the probabilities. That is, the optimization process can effectively find an accurate solution by initializing HMMs with SIGs.

4. Experiment and Evaluation

4.1 Data Description

For the experiment, we use Health Care Reform (HCR) dataset. The HCR dataset was built in March 2010 by crawling tweets containing the hashtag “#hcr” (health care reform). The dataset consists of 839 tweets for the training set, 838 tweets for the development set, and 839 tweets for the test set. Because we use tweets with positive or negative polarity, the training set and the test set have 621 and 665 tweets, respectively.

4.2 Data Preprocessing

Although usernames, URLs, and emoticons are important components of microblogs, they are very diverse. In most cases, we do not need to know their specific values for sentiment analysis. To reduce the feature space for effective classification, we replace them as follows:

Usernames Users often include a username to direct their messages. Usually usernames follow the symbol @. Instead of treating each username as a unigram, we replace them with ‘@user,’ which indicates that a username is here.

Usage of links Because of the limited length, people often include URLs to provide sufficient information in microblogs. We convert URLs such as “http://bit.ly/9Klc8V,” into the token ‘URL.’

Emoticons People often use emoticons in microblogs to directly express their sentiment. All we need is the polarity of the emoticons, so we replace all positive emoticons with

Table 1. Performance comparison with the baselines and the previous work

Approach	Acc.	Positive			Negative			Average		
		P	R	F1	P	R	F1	P	R	F1
Proposed approach ($h = 2$)	<u>0.786</u>	0.535	<u>0.591</u>	<u>0.562</u>	<u>0.873</u>	0.845	0.859	<u>0.704</u>	<u>0.718</u>	<u>0.710</u>
Proposed approach ($h = 3$)	0.782	0.527	0.571	0.548	0.867	0.845	0.856	0.697	0.708	0.702
Proposed approach ($h = 4$)	<u>0.794</u>	<u>0.554</u>	0.565	<u>0.559</u>	<u>0.868</u>	0.863	<u>0.866</u>	<u>0.711</u>	<u>0.714</u>	<u>0.713</u>
SVM	0.719	0.395	0.403	0.399	0.819	0.814	0.816	0.607	0.608	0.608
NB	0.752	0.465	0.481	0.473	0.842	0.834	0.838	0.654	0.657	0.655
HMM	0.779	0.523	<u>0.571</u>	0.545	0.867	0.841	0.854	0.695	0.706	0.699
Speriosu et al. [7]	0.712	-	-	-	-	-	-	-	-	-
Saif et al. [20]	0.682	<u>0.538</u>	0.472	0.503	0.845	<u>0.876</u>	0.860	0.652	0.674	0.682
Silva et al. [21]	0.784	0.529	0.299	0.382	0.813	<u>0.920</u>	0.863	0.671	0.610	0.623

‘;)’ and all negative emoticons with ‘;)’ based on information about commonly used emoticons(http://en.wikipedia.org/wiki/List_of_emoticons).

In the feature selection step, we choose top-10% words building sentiment models.

4.3 Experimental Result

In this section, we present and discuss the experimental result of our proposed approach. Table 1 shows the performance comparison with baselines and the previous approaches.

We implement sentiment classifiers with well-known machine learning algorithms as baselines: SVM, NB, and HMMs which do not use SIGs but randomly initialize hidden states. We set the number of hidden states in the HMMs to 2 with the best performance.

For the comparison with the existing approaches, we choose 3 approaches. If several performance values were presented in the papers, we use the best ones for the comparison.

We present the performance of our methods with different numbers of hidden states (h): 2, 3 and 4. As the performance measures, we use accuracy, precision, recall and F1-measure. Precision, recall and F1-measure are measured for positive sentiment and negative sentiment, and the averages of the two are also presented. In Table 2, underlined bold numbers are the best and bold numbers are the second best. Dashes indicate no reported performance results.

The proposed approach outperforms the baselines and the previous approaches regardless of the number of hidden states. In most of the measures, the top two ranked approaches are ours.

Our approach yields the best performances in most evaluation measures. The proposed approach with $h = 4$ shows higher accuracies, from 1.3% to 16.4%. Especially, the dataset has a class imbalance problem, the approaches by Saif et al. [20] and de Silva et al. [21] showed poor performance of recall in the minority class, Positive. However, the proposed approach reports relatively good performances of recall in the minority class, providing more evidence that SIGs are an effective way to train HMMs.

Among the simple baselines, the performance of the sequential models (HMMs) is impressive. They show better results than non-sequential models (SVMs and NBs) and even comparable to the existing approaches in most of the cases. It can be said that considering word sequences were very effective for the performance improvement of sentiment analysis. Compared with the simple HMMs, the proposed approach produces higher accuracy, it implies shows SIGs very effectively worked.

5. Conclusion

In this paper, we have presented a new sentiment analysis method for microblogs that models the transition patterns of sentimental sentences. Even though sentences in microblogs are sequences of words, most previous research focused on extending feature sets without considering word sequences.

The proposed sentiment analysis approach learns patterns of syntactic and sentimental word transitions. We identified SIGs and used them to initialize HMMs. Because SIGs can be regarded as formative elements of sentimental sentences, using them allows HMMs to more properly model the transition patterns. SIGs help HMMs establish initialization points in an

appropriate range, which can have a positive effect on finding an accurate solution. Thus, our proposed approach was robust with respect to the random initialization seed.

We provided an experimental demonstration and various comparisons with existing works to evaluate the performance of our proposed approach. The results show that our approach outperformed existing approaches.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgment

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. R7120-17-1016, Development of Industry Evaluation Analysis SW based on Convergence of Structured and Unstructured Big Data to provide industry analysis information in a timely manner). Also, this work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2016R1A2B4015820).

References

- [1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in *Proceedings of the Workshop on Language in Social Media (LSM)*, Portland, OR, 2011, pp. 30-38.
- [2] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford University, Technical Report, 2009.
- [3] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: the good the bad and the OMG!," in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, Barcelona, Spain, 2011, pp. 538-541.
- [4] F. Bravo-Marquez, M. Mendoza, and B. Poblete, "Combining strengths, emotions and polarities for boosting Twitter sentiment analysis," in *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, Chicago, IL, 2013, pp. 1-9. <https://doi.org/10.1145/2502069.2502071>
- [5] C. E. Rasmussen, "The infinite Gaussian mixture model," *Advances in Neural Information Processing Systems*, vol. 12, pp. 554-560, 2000.
- [6] H. Saif, Y. He, and H. Alani, "Alleviating data sparsity for twitter sentiment analysis," in *Proceedings of CEUR Workshop*, Lion, France, 2012, pp. 2-9.
- [7] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," in *Proceedings of the 1st Workshop on Unsupervised Learning in NLP*, Edinburgh, Scotland, 2011, pp. 53-63.
- [8] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, 2002, pp. 79-86. <https://doi.org/10.3115/1118693.1118704>
- [9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135, 2008. <http://doi.org/10.1561/15000000011>
- [10] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, et al., "OpinionFinder: a system for subjectivity analysis," in *Proceedings of HIT/EMNLP on Interactive Demonstrations*, Vancouver, Canada, 2005, pp. 34-35. <https://doi.org/10.3115/1225733.1225751>
- [11] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov Models*. Heidelberg: Springer, 1994.
- [12] L. Rabiner and B. H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4-6, 1986.
- [13] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. Boston, MA: Springer, 2009, pp. 827-832. https://doi.org/10.1007/978-1-4899-7488-4_196
- [14] A. Bifet and E. Frank, "Sentiment knowledge discovery in twitter streaming data," in *Discovery Science*. Heidelberg: Springer, 2010, pp. 1-15. https://doi.org/10.1007/978-3-642-16184-1_1
- [15] C. Hutto and E. Gilbert, "Vader: a parsimonious rule-based model for sentiment analysis of social media text,"

- in *Proceedings of 8th International AAAI Conference on Weblogs and Social Media*, Ann Arbor, MI, 2014.
- [16] K. Liu, W. Li, and M. Guo, "Emoticon smoothed language models for twitter sentiment analysis," in *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, Toronto, Canada, 2012, pp. 1678-1684.
- [17] S. Matsumoto, H. Takamura, and M. Okumura, "Sentiment classification using word sub-sequences and dependency sub-trees," in *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Hanoi, Vietnam, 2005, pp. 301-311. https://doi.org/10.1007/11430919_37
- [18] S. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: building the state-of-the-art in sentiment analysis of tweets," in *Proceedings of 2nd Joint Conference on Lexical and Computational Semantics*, Atlanta, GA, 2013, pp. 321-327.
- [19] H. Saif, M. Fernandez, Y. He, and H. Alani, "Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the STS-Gold," in *Proceedings of 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM)*, Turin, Italy, 2013.
- [20] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of Twitter," in *The Semantic Web-ISWC 2012*. Heidelberg: Springer, 2012, pp. 508-524. https://doi.org/10.1007/978-3-642-35176-1_32
- [21] N. F. da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, pp. 170-179, 2014. <https://doi.org/10.1016/j.dss.2014.07.003>
- [22] D. Ziegelmayer and R. Schrader, "Sentiment polarity classification using statistical data compression models," in *Proceedings of IEEE 12th International Conference on Data Mining Workshops*, Brussels, Belgium, 2012, pp. 731-738. <https://doi.org/10.1109/ICDMW.2012.43>
- [23] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," *LREC*, vol. 10, pp. 2200-2204, 2010.
- [24] L. Polanyi and A. Zaenen, "Contextual valence shifters," in *Computing Attitude and Affect in Text: Theory and Applications*. Dordrecht: Springer, 2006, pp. 1-10. https://doi.org/10.1007/1-4020-4102-0_1

Noo-ri Kim received the B.S. degree in computer engineering from Sungkyunkwan University, Suwon, Korea in 2013. He is currently pursuing his M.S./Ph.D. in Computer Engineering at Sungkyunkwan University. His research interests include recommender systems, text mining, and machine learning.

E-mail: pd99j@skku.edu

Kyoungmin Kim received her B.S. in medical computer science from Eulji University, Seongnam, Korea, in 2013. She received her M.S. in computer engineering from Sungkyunkwan University, Suwon, Korea, in 2015. Her research interests include text mining and sentiment analysis.

E-mail: kmkim1222@skku.edu

Jee-Hyong Lee received his B.S., M.S., and Ph.D. in computer science from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea, in 1993, 1995, and 1999, respectively. From 2000 to 2002, he was an international fellow at SRI International, USA. He joined Sungkyunkwan University, Suwon, Korea, as a faculty member in 2002. His research interests include text mining, intelligent systems, and machine learning.

E-mail: john@skku.edu