Data & Knowledge Engineering 110 (2017) 54–67

Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak

An adaptable fine-grained sentiment analysis for summarization of multiple short online reviews

Reinald Kim Amplayo, Min Song*

Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, Republic of Korea

ARTICLE INFO

Keywords:
Review summarization
Aspect extraction
Sentiment analysis
Short texts
Online reviews

ABSTRACT

In this study, we present a novel method in generating summaries of multiple online reviews using a fine-grained sentiment extraction model for short texts, which is adaptable to different domains and languages. Adaptability of a model is defined as its ability to be easily modified and be usable on different domains and languages. This is important because of the diversity of domains and languages available. The fine-grained sentiment extraction model is divided into two methods: sentiment classification and aspect extraction. The sentiment classifier is built using a three-level classification approach, while the aspect extractor is built using extended biterm topic model (eBTM), an extension of LDA topic model for short texts. Overall, results show that the sentiment classifier outperforms baseline models and industry-standard classifiers while the aspect extractor outperforms other topic models in terms of aspect diversity and aspect extracting power. In addition, using the Naver movies dataset, we show that online review summarization can be effectively constructed using the proposed methods by comparing the results of our method and the results of a movie awards ceremony.

1. Introduction

The Internet has become a pivotal channel of communication and interaction between online consumers and producers. Social media where the communication between two parties take place have rich but unorganized content contributed by users, often in fragmented and sparse fashion. Normally both consumers and producers refer to online reviews for several reasons [1]. Consumers look at them to decide whether to purchase the presented product/service or not based on their demands. Producers, on the other hand, look at them to improve their market strategy by magnifying the positive aspects and improving the negative ones. Vermeulen et al. [1] and Ye et al. [2] showed that the existence of online hotel reviews enhances hotel consideration in consumers and business performance of hotels. Duan et al. [3] also showed with the use of online movie reviews as data that the volume of online reviews significantly influences the sales. Chatterjee [4] also looked at the influence of negative consumer reviews on retailer evaluation and patronage intention.

Meanwhile, although the amount of information is constantly increasing, information available on the Internet is slowly becoming shorter. The most popular example of this trend is Twitter. Also, there are a few and growing number of online review websites that follow Twitter's short character limit. Both the biggest movie review websites in South Korea, Naver movie and Daum movie, limit user reviews to 140 and 150, respectively. This raises a concern to the sentiment analysis problem because limiting the number of characters limits the number of words that can be inputted for the review. This in turn changes the natural language properties of the text; syntactical errors are more common, the context is direct, and the aspect being reviewed is limited to mostly one aspect [27].

* Corresponding author.
E-mail addresses: rktamplayo@yonsei.ac.kr (R.K. Amplayo), min.song@yonsei.ac.kr (M. Song).

http://dx.doi.org/10.1016/j.datak.2017.03.009
Received 17 August 2016; Received in revised form 2 March 2017; Accepted 17 March 2017
Available online 03 May 2017

One primary concern with sentiment analysis of movie review is to automate the comprehension of the vast amounts of online reviews. One way to solve this is to create a review summary that provides users a condensed outline of list of aspects and their corresponding sentiment rating. There are considerable efforts on fine-grained sentiment extraction, a lot of which uses topic modeling techniques [5–8] to automate the extraction of aspects. But with the quickly growing amounts of short texts data, there is an urgent call to designing a fine-grained sentiment extraction model specifically for short texts.

The purpose of this study is to propose an adaptable approach to creating a review summarization framework for analyzing short social media text with novel sentiment and aspect extraction algorithms. Past research works focused on the flexibility of a model. Flexibility is defined as the model's ability to be independent in terms of domain, topic, temporal, and language style [9]. On the other hand, in this paper, we aim at developing the adaptable approach for summarization of short multiple reviews. We define adaptability as the ability of an approach to be easily modified for the creation of models for other domains or languages. Notice that *models* are flexible and *approaches* are adaptable. An approach is used to create a model. There can be multiple products from different domains as well as reviews written using different kinds of languages. Creating different approaches for these different domains and languages takes a considerable amount of effort and time. Thus, we attempt to achieve the adaptability in our models with a corpus-driven approach, where external data other than the given corpus are not used.

The proposed approach is divided into two parts: sentiment classification and aspect extraction. The first part is the building of a three-level sentiment classifier. The first level consists of classifiers using natural language processing techniques, specifically lexical and syntactical techniques. The second level classifiers are two support vector machines classifiers, handling different n-gram feature vectors from different dictionaries. The last level combines the first two levels using a simple feedforward neural network. The second part is the construction of an aspect extraction model. The aspect extractor is built using extended biterm topic model (eBTM), which inserts a document layer normally found on LDA [10] outside the biterm layer of the original BTM [11] to cope with the weaknesses of both models on short and narrow-scoped texts. For emphasis, we do not use external data other than the corpus for adaptability.

Several experiments are carried out using multiple datasets from multiple languages and multiple domains, in order to verify the adaptability of the model. The results can be summarized as follows. The proposed sentiment classifier outperforms previous state-of-the-art classifiers and topic model-based classifiers. The proposed aspect extraction method outperforms LDA [10] and BTM [11] in terms of aspect diversity score and aspect extracting power. Finally, we show that the proposed architecture combining the sentiment classification and aspect extraction approach is effective in summarizing reviews, based on a case study on Korean movies.

The rest of the paper is organized as follows: the related work section discusses some previous research on adaptable sentiment analysis and fine-grained sentiment extraction, and topic models for short texts. The methodology section describes our datasets and the preprocessing stage, and the construction of our sentiment classification and aspect extraction models. The results section reports the experimental results and the evaluation of our methods compared to other methods. Finally, the conclusion section summarizes the contributions of our work and present future directions.

2. Related work

Since one of the aims of our study is to develop adaptable sentiment and aspect extraction, we conducted a literature review on related studies to the present paper.

2.1. Adaptable sentiment analysis for short texts

A lot of recent literature on sentiment analysis deal with flexibility, more explicitly with the dependency and independency on both domain and language. Inflexibility, or the dependency on a domain of a model is used to improve the performance of the domain-specific classifier [12,13]. Domain independency can be achieved by the use of a large collection of unlabeled data from different domains [14,15], while language independency can be achieved by using no additional linguistic information nor external resources [16].

In comparison with the number of research done on flexibility, there are quite a few studies that target adaptability on sentiment analysis. Sentimatrix [17] combines rule-based classification, statistical and machine learning methods in creating a sentiment analysis technique. They stated that grammar-based systems typically obtain better precision but are hard to adapt to new domains. Therefore, they designed the technique to be language independent, only using resources that are easily adaptable for any language such as tokenizers and annotations. Tanev et al. [18] also mentioned the advantage in terms of adaptability of using a bag of n-gram representations because it is easily adaptable to any languages.

There are various approaches to sentiment analysis that target short texts using machine learning [19–21] and natural language processing [22–24]. One of the most recent and popular works is carried out by Yang et al. [21], where they incorporated several syntactic features including POS tags and n-grams selected by a measure of information gain into computing sentiment similarity for Chinese microblogs. Another recent work is proposed by Thelwall et al. [22], where they proposed SentiStrength for MySpace comments, which uses human-annotated sentiment scores of computing the word sentiment strength. This was adopted by Basiri et al. [23] where they added the comment history of the reviewers to improve the sentiment score. Another line of research is to combine machine learning and natural language processing for sentiment analysis. Aldayel et al. [25] use both lexical-based and the support vector machines classifier to for hybrid sentiment analysis for Arabic tweets.

Just a few studies on adaptable sentiment analysis for short texts are available. Balahur [26] designed a sentiment analysis tool that works with Twitter data using minimal linguistic processing (only used tokenizers) which makes the approach easily portable to other languages. They also mentioned that to benefit from Twitter's multi language data, the system should be adaptable to other languages as well.

2.2. Fine-grained sentiment extraction

In order to make use of sentiment analysis and create a review summarization, a fine-grained sentiment extraction must be constructed, where aspects must also be incorporated. Methods used for fine-grained sentiment extraction can be divided into two: non-topic modeling techniques and topic modeling techniques.

Non-topic modeling techniques make use of natural language processing techniques more compared to topic modeling techniques. These techniques usually require external resources, such as knowledge bases or human annotators. Thelwall et al. [27] used traditional approaches to creating annotation; three independent human coders created a standard term/aspect list and also annotated the Twitter data manually. Jimenez-Zafra et al. [28] used Freebase, which contains information about different domains, to extract the aspects and combined different linguistic resources and SentiWordNet to classify sentiments. On the other hand, Miao et al. [29] created their own knowledge bases for aspects and sentiment words using both semi-structured and unstructured reviews. Thet et al. [30] used an approach to computing the sentiment, by assigning them to individual words and considering grammatical dependency to compute the sentiments of movie reviews on discussion boards. They used a separate external list of aspects and connected these aspects to the reviews.

Topic modeling techniques make use of latent Dirichlet allocation (LDA) or its variants to automatically extract aspects from text. Bagheri et al. [8] used an extension of LDA that can extract multiword aspects from text collections. However, unlike our approach presented in this paper, they did not incorporate these aspects into a sentiment classifier to build a fine-grained sentiment extraction framework. Titov et al. [5] constructed a joint model of text and aspect ratings for sentiment summarization. They provided their own predefined aspects and feed them to their model. Jo et al. [6] also constructed a joint model called aspect and sentiment unification model (ASUM), but instead of creating their own predefined aspects, they used external sentiment seed words to help with the classification of sentiment. They assume that all words in one sentence only refer to one aspect. This is immediately extended by Kim et al. [7] which makes use of a tree-structured Bayesian nonparametrics method called recursive Chinese Restaurant Process (rCRP) to also extract the aspect's hierarchy.

2.3. Topic modeling for short texts

Since LDA was proposed in Blei et al. [10], it has gained a great attention in the research community of social media where texts to be processed is relatively short. Karandikar et al. [31] used LDA to cluster short status messages. Godin et al. [32] used LDA for recommending hashtags for tweets without hashtags. LDA is a very effective way to extract topics from documents, but it is not as effective when dealing with short texts, as short texts are very sparse. Because of this, various studies tackle the problem of constructing a topic model specifically for short texts.

One way to cope up with this shortcoming is to change the input documents in some way. Kim et al. [33] used LDA with n-grams as its vocabulary to increase the volume of input data. Zhu et al. [34] improved the topic model by introducing an external knowledge base to identify topics from text for better categorization. They also added weight to few features in the texts to get some other topics from the topic model. Hong et al. [35] trained the topic model using Twitter messages aggregated by the one who created the message. This made the quality of learned topic even better than the Author-Topic model, which fails to model hierarchical relation between entities.

Another novel way to cope up with the weakness of LDA is to extend the model by adding layers and variables. Xu et al. [36] developed Twitter-user model as an extension of Author Topic model to discover Twitter user's interest. They introduced a latent variable that indicates whether it is related to the author's interest or not. Yan et al. [11] extended the unigram mixture model and constructed a biterm topic model for short text. They posited that instead of relying on the implicit pattern connection of LDA, directly modeling the generation of word co-occurrence patterns, hereby called biterms, is a better way to model topics for short text. They solved the sparseness problem by using aggregated patterns in the whole corpus for learning topics. Another extension was made by Lin et al. [37] called dual-sparse topic model which mines focused topics and focused terms in short text. They used a method called “spike and slab”, which decouples the sparsity and smoothness of the distributions found in LDA.

Unlike previous works, we focus on the adaptability of our model in terms of domain and language to create a fine-grained sentiment extraction framework. This means external resources are not used and the creation of the model is corpus-driven. We incorporate the idea in [25] to combine both natural language processing and machine learning techniques but in a different way. Unlike previous studies on fine-grained sentiment extraction, we primarily focus on short texts as the target dataset. We also extend the ideas and notions on previous studies [10,11] on topic modeling on short texts to improve the extraction of aspects. Finally, we propose a novel method in generating summaries of multiple reviews.

3. Methodology

The main idea of our approach comes from the assumption that since the dataset consists of only short texts, one text may reveal only one aspect and thus only has one sentiment. This assumption is based from the empirical theory that one sentence contains one aspect [6]. Since the texts are short, we can assume that most of the texts have at most one sentence. From this assumption, we suggest that it is possible to separate the models for sentiment classification and the aspect extraction to cater a better performance

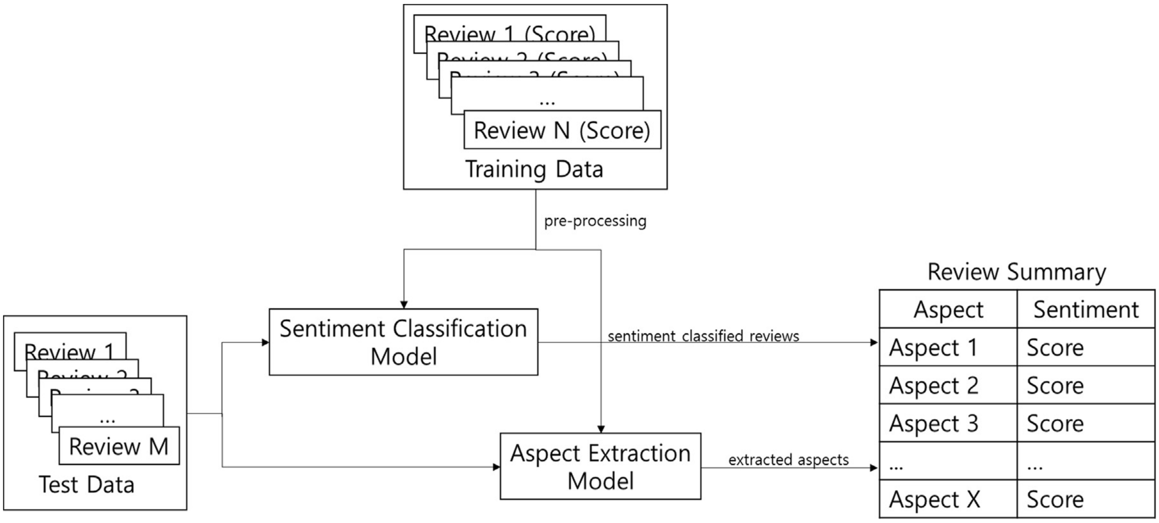


Fig. 1. Overview of the review summarization framework. A review summary is generated by combining the results of the sentiment classification model (i.e. sentiment score), and of the aspect extraction model (i.e. aspect categories).

and to consider the mode's adaptability. The proposed review summarization framework is shown in Fig. 1. After preprocessing, a separate training data is used to train the sentiment classification and the aspect extraction model. Using separate review datasets that are not yet classified, the review summary is generated by going through the constructed sentiment classification model and aspect extraction model.

In the methodology section, we present the two major parts of the summarization framework of multiple short reviews: the first one is about the sentiment classification model and the second one is about the aspect extraction model.

3.1. Datasets

We collect multiple datasets to accurately evaluate adaptability of our approach. After collecting data, we filter those data whose character length exceeds 140. Data from two different domains, movies and products, are gathered from English websites Rotten Tomatoes and Amazon computer products, respectively. Correspondingly, movie reviews data from languages other than English are gathered from Naver movie for Korean data and from Douban for Chinese data. We limit the collection of data to 10,000 movie reviews. We divide the data into train and test data. Table 1 shows the basic statistics of the collected data. In order to show the frequencies of different data comparatively with each other, we normalize them to fit a [0, 1] range in a range between the real numbers 0 and 1, inclusive.

Text preprocessing is carried out in the following steps. We use the simple tokenizer and the POS tagger for three languages used in this paper: Korean, Chinese, and English. The Korean tokenizer and the POS tagger called twitter-korean-text[1] is used to tokenize and POS tag the Korean documents. Stanford CoreNLP [38] is used to tokenize, lemmatize, and POS tag the English documents and the Chinese documents. We do not use stopword removal as this requires an external list which might not be available on some other languages, thus hurts the model's adaptability. We also extract separately the list of nouns of each document. This is used for the aspect extraction model construction.

Table 1
Statistics of datasets.

Domain	Movie			Product
Language	Korean	Chinese	English	English
Website	Naver	Douban	Rotten Tomatoes	Amazon
#comments	10000	10000	10000	10000
#1.0 rating	4078	2013	3179	6895
#0.9 rating	1026	–	1580	–
#0.8 rating	1056	3862	1932	1710
#0.7 rating	744	–	875	–
#0.6 rating	556	2905	721	441
#0.5 rating	496	–	357	–
#0.4 rating	299	818	370	389
#0.3 rating	212	–	156	–
#0.2 rating	243	402	350	565
#0.1 rating	1290	–	480	–

We build several dictionaries and lists without using external resources. We construct noun and verb lists by getting the top 1000 words by frequency from the training data. We also construct four other lists: good and bad adjective lists, and good and bad adverb lists, by getting the top 500 words of each parts of speech by frequency from the training data. We also make two n-gram dictionaries: character-based n-gram dictionary and word-based n-gram dictionary. The character n-gram dictionary accepts the original text as input while the word n-gram dictionary accepts the list of tokenized strings as input. The n-gram dictionaries are created with an established n (character: 4 to 8, word: 1 to 3) and reduced by removing the less frequent n-grams.

3.2. Multi-level classification

The sentiment classification model proposed in this paper is a multi-level classification model where the first two levels use two widely used approaches in sentiment classification: natural language processing techniques and supervised learning techniques. The third and final level combines the first two levels using a simple neural network algorithm. The tokens produced along with the dictionaries and lists built in the preprocessing stage are used to extract the features from texts.

3.2.1. First level: natural language processing

With the use of the constructed noun and verb lists, and the extracted good/bad adjective/adverb lists (Section 3.1), we calculate naïve sentiment scores based on natural language processing techniques. There are four lexical-based scores and two syntactic-based scores. The lexical-based scores are calculated using the good/bad adjective/adverb lists. These lists are sorted by frequency and thus can be interpreted as sentiment weights; the higher the frequency of the word, the more it is inclined to the polarity of the list. Based on the lists, we then create a lexical score function that gives a score to a word, as shown below. This function gives a score to a word based on its rank (position in the list).

$$\mathrm{LS}(\mathrm{word},\ \mathrm{list}) = \begin{cases} 2^{\frac{len(list)}{100}}, & if\ rank(word,\ list) < 50 \\ 2^{\frac{len(list)}{100}-n}, & if\ rank(word,\ list) < n*100 \end{cases}$$

The function gives a bigger score to a word that has a higher ranking compared to a word that has a lower ranking in the list. For example, in a 500-word list, the first 50 words get a score of 32 (2^5), the next 100 words get a score of 16 (2^4) and so on. Using this function, we now create four different lexical scores by applying the above function for each word in the document with each available list.

We also calculate two syntactic-based scores which are called naïve lazy syntactic scores. These scores are based on the assumption that there are two possible syntaxes of a sentence with a sentiment. The first one is the noun-adjective syntax where there is an adjective describing the noun. Another syntax is the noun-adverb-verb syntax where there is an adverb describing the action of the noun given by the verb. For example, the sentence “the actor was great” follows the noun-adjective syntax where actor is the noun and great is the adjective. Another example is the sentence “the actor acts well” where it follows the noun-adverb-verb syntax where actor is the noun, acts is the verb, and well is the adverb. Note that the order of these words does not matter. We make use of the lexical score function above to add more weight on those words with high frequency. We normalize these weights by dividing them to the distance between the words. In the case of adjectives and adverbs, we first check whether the said adjective or adverb is in the bad or the good list. The functions for the noun-adjective syntax and the noun-adverb-verb syntax are shown below.

$$SS(n,\ adj) = \frac{LS(n,\ nouns)*LS(adj,\ [bad,\ good]adjectives)}{distance(n,\ adj)}$$

$$SS(n,\ adv,\ v) = \frac{LS(n,\ nouns)*LS(adv,\ [bad,\ good]adverbs)*LS(v,\ \ verbs)}{\max\ (distance(n,\ adv),\ distance(n,\ v),\ distance(adv,\ v))}$$

3.2.2. Second level: machine learning

In this stage, we use the character-based n-gram dictionary and the word-based n-gram dictionary to build two sparse matrices. These matrices are then fed to two different support vector machine classifiers. We use an L2-loss L2-regularized support vector regression provided by LIBLINEAR [39], a library for large linear classification specifically good for document classification. Each support vector machine classifier receives a sparse matrix as an input and the sentiment score as the output and produces a real number between 0 and 1, inclusive.

The role of the second level classification is to look at the patterns created by continuous characters and words. More specifically, the character-based support vector machine classifier takes care of continuous character patterns in the text. Short texts found in the web are most likely to be dirty. Character-based support vector machines classifier tackles that problem. Meanwhile, word-based support vector machines classifier accepts a cleaned and tokenized set of strings. Therefore, word-based support vector machines classifier classifies the text after it is cleaned.

3.2.3. Third level: neural network

The final level of the classification model is the combination of the first two levels. This is accomplished using the classical feedforward neural network and the backpropagation algorithm. We combine all the outputs from the past levels, the lexical and

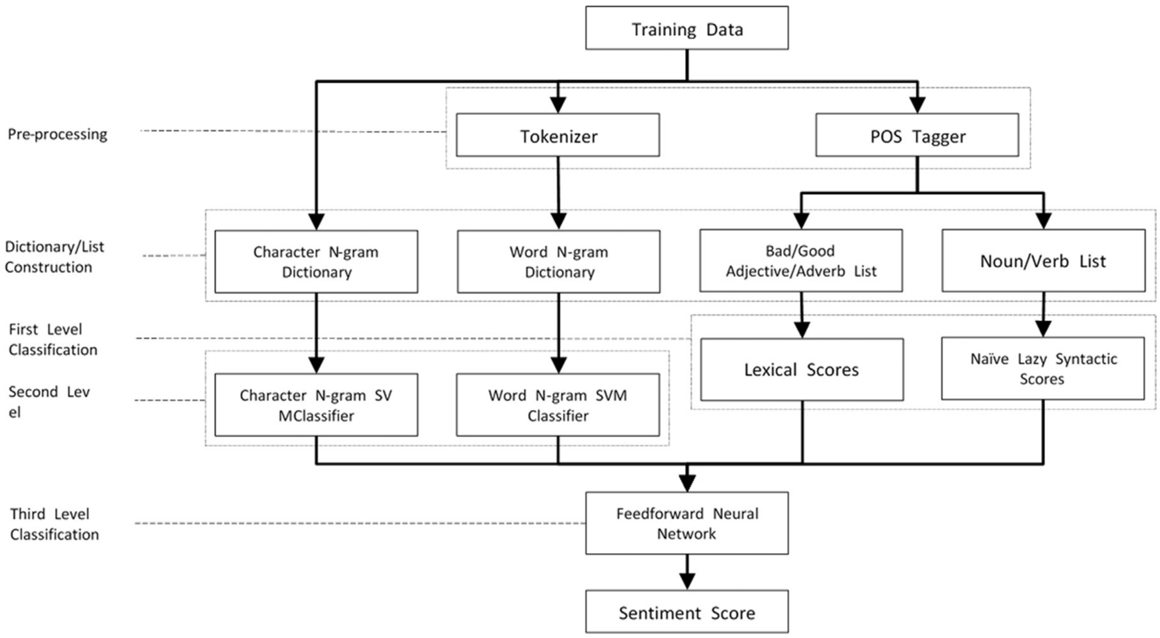


Fig. 2. Sentiment three-level classification model framework.

syntactic scores from the first level, and the outputs of the support vector machine classifiers from the second level, for a total of 8 features and feed all of them as an input of the neural network. The hidden layer consists of 15 nodes and the output layer consists of a single node that contains the sentiment score. The final sentiment score is a real number between 0 and 1, inclusive.

The whole system is shown figuratively in Fig. 2.

3.3. Extended biterm topic model

The aspect extraction proposed in this paper is a modification of the existing biterm topic model, which we call extended biterm topic model (eBTM). The main difference between LDA and BTM is that LDA implicitly considers the word co-occurrence patterns inside a document by adding a document layer. However, because of the sparse word co-occurrence patterns on short text, the implicit patterns become less effective. BTM suggests that if the co-occurrence patterns are explicitly generated, i.e. using biterms, they can become more effective for short texts. Biterms can be simply described such that in the short text "biterm topic model", the extracted biterms are "biterm topic", "topic model", and "biterm model". A biterm does not have a word order, i.e. "biterm topic" and "topic biterm" are the same.

But because BTM explicitly patterns out word co-occurrence, the whole collection of documents only has one shared topic distribution. We extend this approach by incorporating the document layer of LDA. eBTM turns back to LDA to learn the co-occurrence patterns of words implicitly using the document layer and uses the biterm approach of BTM to expand the size of the short text. This way, eBTM still considers the word co-occurrence patterns while at the same time creates a document level topic distribution. Fig. 3 shows the difference between the three topic models.

3.3.1. eBTM generative process

Since eBTM is derived from both LDA and BTM, consequently its generative process is also similar to them. The specific generative process for a document collection D under the eBTM model is as follows:

1. For each topic $k \in K$:
    1.1 $\varphi^{(k)}$~Dirichlet(β)

2. For each document $d \in D$:
    2.1 θd~Dirichlet(α)
    2.2 For each biterm $b \in B$:
        2.2.1 $z \sim Discrete(\theta d)$
        2.2.2 $wi \sim Discrete(\varphi^{(z)})$
        2.2.3 wj~$Discrete(\varphi^{(z)})$

where K is the number of latent topics in the collection, $\varphi^{(k)}$ is a topic-specific discrete probability distribution over a fixed vocabulary that represents the kth topic distribution, θ_d is a document-specific discrete probability distribution over the fixed

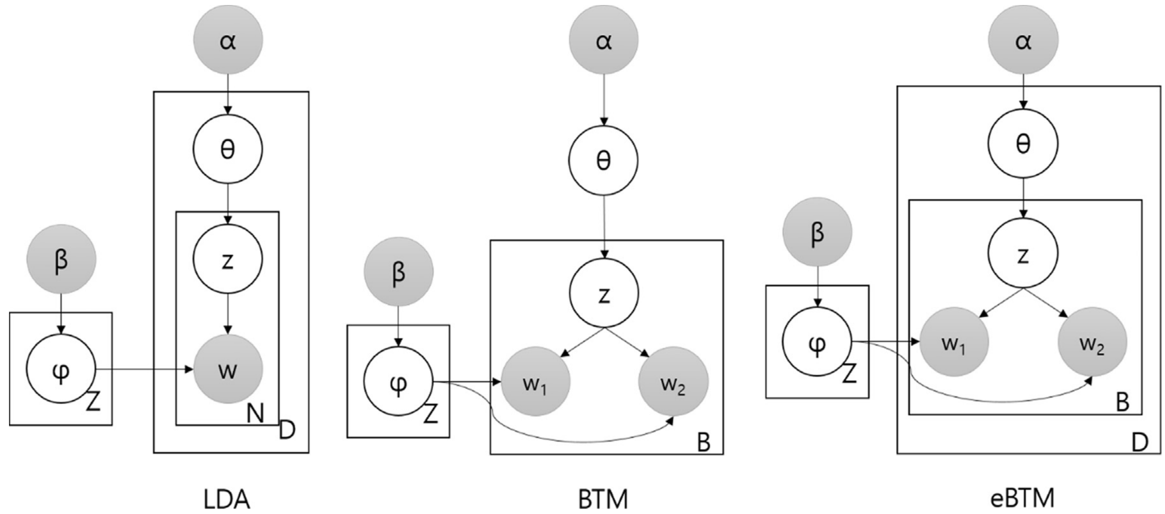


Fig. 3. From left to right: the general topic model latent Dirichlet allocation model, the biterm topic model for short text, and our proposed extended biterm topic model.

number of topics, z is the topic index for the current biterm b, and w_i and w_j are the terms of the current biterm b. The hyperparameters α and β are the same symmetric Dirichlet distributions of LDA where the discrete distributions are drawn from.

3.3.2. Parameters estimation

This section describes the Gibbs sampling method to estimate the latent variables θ and φ. At each transition step of the Markov chain, the topic z is drawn from the conditional probability

$$\mathrm{P}(zZ_{-b},\, B,\, \alpha,\, \beta)\propto(n_{d,k}^{(-b)}+\alpha_k)\frac{n_{k,w}^{(-i)}+\beta_w}{\sum_{w'} n_{k,w'}^{(-i,-j)}+\beta_{w'}}\frac{n_{k,w}^{(-j)}+\beta_w}{\sum_{w'} n_{k,w'}^{(-i,-j)}+1+\beta_{w'}} \tag{1}$$

where Z is the set of topics, B is the set of all biterms, $n_{d,k}$ is the number of biterms assigned to topic k in document d, $n_{k,w}$ is the number of times word w is assigned to topic k, and a minus subscript or superscript indicates the exception of word/s i and/or j, or the biterm b in the count.

With the counters $n_{d,k}$ and $n_{k,w}$, we can easily estimate the topic-word distribution φ and the document-topic distribution θ as:

$$\varphi_{k,w} = \frac{n_{k,w} + \beta}{\sum_w n_{k,w} + M\beta} \tag{2}$$

$$\theta_{d,k} = \frac{n_{d,k} + \alpha}{\sum_k n_{d,k} + K\alpha} \tag{3}$$

where *M* is the number of vocabulary and *K* is the number of topics.

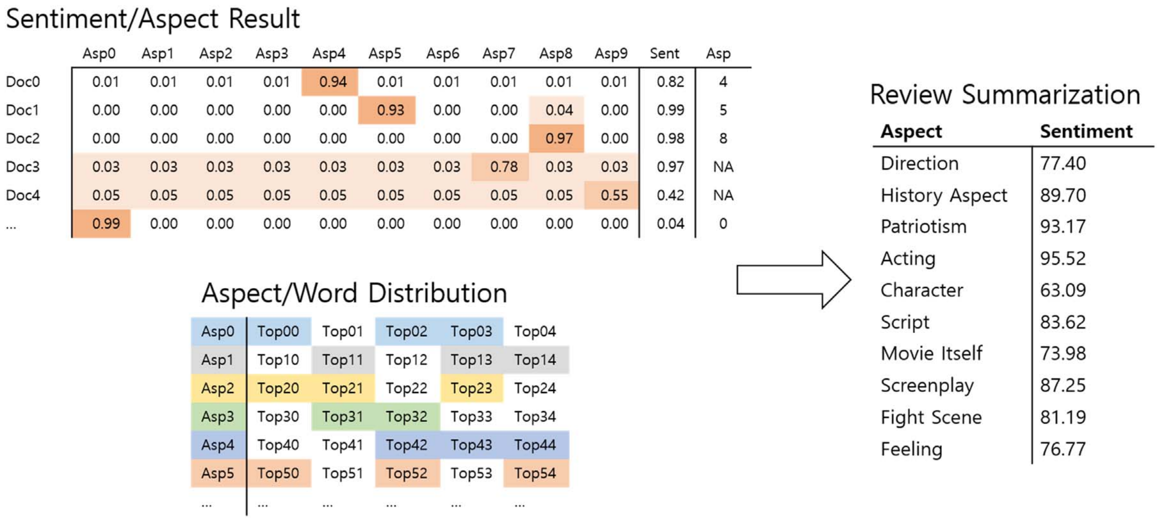
3.4. Review summarization

The main purpose of this paper is to effectively develop a review summarization technique for short texts. We incorporate the built sentiment classification and aspect extraction models to the construction of a review summarization technique. This is achieved by the use of the sentiment-classified and aspect-extracted text, and the aspect-word distribution generated by the aspect extraction model. Fig. 4 shows the simplified framework and an example of a review summary of a movie.

4. Results

4.1. Sentiment classification

This section reports on the results of sentiment classification on texts from multiple domains and languages. We first evaluate the necessity of the multi-level classification model. We change the predicted and actual scores into their polarity based on the fifty percent threshold; the sentiment score of 0.5 above is given a positive sentiment and negative sentiment if otherwise. We compare the final results to the accuracies obtained by the first- and second-level classifiers. Each level of classifiers produces multiple classifications. For example, the second-level classifier produces a classification made by the word n-gram SVM classifier and a classification made by the character n-gram SVM classifier. For simplicity, we get the average of the classifications of each level of classifiers and the results are presented in Table 2. Table 2 shows that the combination of the two different techniques with the use of



Sentiment/Aspect Result

	Asp0	Asp1	Asp2	Asp3	Asp4	Asp5	Asp6	Asp7	Asp8	Asp9	Sent	Asp
Doc0	0.01	0.01	0.01	0.01	0.94	0.01	0.01	0.01	0.01	0.01	0.82	4
Doc1	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.00	0.04	0.00	0.99	5
Doc2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.00	0.98	8
Doc3	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.78	0.03	0.03	0.97	NA
Doc4	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.55	0.42	NA
...	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0

Aspect/Word Distribution

Asp0	Top00	Top01	Top02	Top03	Top04
Asp1	Top10	Top11	Top12	Top13	Top14
Asp2	Top20	Top21	Top22	Top23	Top24
Asp3	Top30	Top31	Top32	Top33	Top34
Asp4	Top40	Top41	Top42	Top43	Top44
Asp5	Top50	Top51	Top52	Top53	Top54
...	...	...	...	...	...

Review Summarization

Aspect	Sentiment
Direction	77.40
History Aspect	89.70
Patriotism	93.17
Acting	95.52
Character	63.09
Script	83.62
Movie Itself	73.98
Screenplay	87.25
Fight Scene	81.19
Feeling	76.77

Fig. 4. Review summarization framework.

Table 2
Classification results compared to single-level classification.

	Naver (Korean movie)	Douban (Chinese movie)	Rotten Tomatoes (English movie)	Amazon (English products)
First-level (NLP)	0.39	0.51	0.26	0.10
Second-level (Supervised)	0.57	0.30	0.65	0.90
Final classifier	**0.86**	**0.75**	**0.86**	**0.98**

a neural network technique significantly improves the accuracy of the classification model.

We move on to the evaluation of the whole classification model in comparison with other classifiers. We compare the performance of our classifier to both LingPipe [40] (unigrams and bigrams), Stanford CoreNLP's sentiment classifier [38] and ASUM [6]. LingPipe classifies texts by first separating subjective sentences from objective sentences and then finds sentiments using word features. Stanford CoreNLP created a sentiment classifier based on deep learning technique called recursive neural network that builds on top of grammatical structures. ASUM is a joint aspect-sentiment topic model that automatically finds sentiment of each found aspect with the use of sentiment seed words. Since ASUM needs seed words in order to run perfectly, we are not able to run ASUM on Korean and Chinese texts because of the unavailability of such golden standard list in the said language. This is one of the inadaptability problems when combining the aspect extraction and sentiment classification. The same goes for the CoreNLP classifier because the model only supports English texts. This is another inadaptability problem when using extensive grammatical structures as features to learn a sentiment classification model.

The evaluation results from the Naver dataset, the Douban dataset, the Rotten Tomatoes dataset, and the Amazon dataset are presented in Table 3. We present four kinds of evaluation metrics: precision, recall, f-score and accuracy. In all cases, our classifier outperforms all the other classifiers except for the results from the Naver dataset, where our classifier ties with the bigram LingPipe classifier. Interestingly, ASUM did not perform well as presented in [7]. Since they used longer texts, it might be that the sentiment classification does not work properly when fed with short texts.

4.2. Aspect extraction

In this section, we present the results gathered from the extraction of aspects using eBTM. The assignment of aspects is done manually by looking at the aspect words discovered by the model. We compare our model to LDA, BTM, and ASUM in terms of aspect diversity and aspect extracting power. We use the Naver dataset to evaluate the aspect diversity and compare it to LDA and BTM, and we use the Rotten Tomatoes dataset to evaluate the aspect extracting power and compare it to ASUM. The number of topics of all models are set to 10, except for ASUM where 20 topics should be set because it extracts the positive and negative aspects differently. We use 0.1 as the alpha and 0.01 as the beta hyperparameters of LDA, BTM, and eBTM. We use 0.1 as the alpha and 1 as the gamma hyperparameters of ASUM. The beta hyperparameter of ASUM is set as follows: 0.001 if the word at hand is not a seed word, 0.1 if the word is a seed word of the topic at hand, and 0 if the word is a seed word of another topic.

4.2.1. Aspect extracting power

The aspect extracting power of a topic model tells us the topic model's capability to extract more aspect terms and less non-aspect

Table 3
Classification results compared to other methods.

Dataset	Evaluation	LingPipe (UNI)	LingPipe (BI)	CoreNLP (RNN)	ASUM	Our own
Naver	Precision	90.9	94.3	NA	NA	**95.5**
(Korean movie)	Recall	**89.3**	89.2	NA	NA	88.2
	F-Score	90.1	91.7	NA	NA	**91.7**
	Accuracy	84.0	86.0	NA	NA	**86.0**
Douban	Precision	85.0	91.5	NA	NA	**97.6**
(Chinese movie)	Recall	**79.6**	78.1	NA	NA	76.5
	F-Score	82.2	84.3	NA	NA	**85.8**
	Accuracy	72.0	74.0	NA	NA	**75.0**
Rotten Tomatoes	Precision	**96.2**	88.8	95.1	89.8	90.4
(English movie)	Recall	84.7	90.2	70.0	87.7	**92.4**
	F-Score	90.1	89.5	80.6	88.8	**91.4**
	Accuracy	82.0	83.0	71.9	81.0	**86.0**
Amazon	Precision	95.8	**99.3**	98.3	70.1	99.2
(English product)	Recall	92.3	97.8	64.2	96.8	**98.1**
	F-Score	94.0	98.6	77.6	81.3	**98.7**
	Accuracy	89.0	97.0	66.6	71.0	**98.0**

terms. Therefore, to evaluate a topic model's extracting power, we need to determine the average number of extracted aspect terms per aspect. This can be done in the following simple manner. We first remove the non-meaningful terms. We then count the number of aspect terms and get the average. We also want to check the average number of unique aspect terms, so we remove the duplicates and get the average.

We reuse the data in Table 4 to compare the aspect extracting power of eBTM with the aspect extracting power of LDA and BTM. We also run eBTM and ASUM on the Rotten Tomatoes data and compare the different model's aspect extracting power. Results are shown in Table 5. Using the Naver dataset, eBTM still outperforms both LDA and BTM in terms of aspect extracting power; eBTM has 2.2 aspect terms on average, 1.3 when duplicates are removed while BTM has 2.1 aspect terms on average, 1.2 when duplicates are removed, and LDA has 1.8 aspect terms on average, 1.3 when duplicates are removed. Using the Rotten Tomatoes dataset, eBTM clearly outperforms ASUM, extracting 2.5 aspect terms on average, 1.9 when duplicates are removed, compared to 1.05 extracted aspect terms, 0.6 when duplicates are removed.

4.2.2. Aspect diversity score

The aspect diversity score checks how diverse the extracted aspects are. Therefore, to check whether a model has more diverse aspects than another model, we need to count the number of certain and valid extracted aspects of the model. To do this, the diversity score is calculated as follows. We gather the first five terms extracted by each aspect. We then remove non-meaningful terms and mark the remaining terms their aspect types. Non-meaningful terms are terms that are included in the first few terms of the topic but have not aspect type. For example, the word “movie” and “film” are considered non-meaningful terms. We then mark aspects with all terms non-meaningful as no aspect. We also mark aspects with more than two aspect types as ambiguous. We then mark aspects with only one meaningful term as weak. We count the number of remaining aspects and count the number of unique aspects.

There are seven types of aspects extracted by the three different topic models. These are presented in Table 4. The aspects extracted by each topic model and the corresponding markings are presented in Table 6. The results show that eBTM outperforms both LDA and BTM in terms of aspect diversity score; eBTM extracted five aspects, three of them are unique while BTM extracted only three aspects, two of them are unique and LDA extracted only two unique aspects. All data are translated from Korean to English.

4.3. Review summarization

We use the data in the Blue Dragon Film Awards 2014[2] to look for possible movie comparisons in order to check the effectivity of

Table 4
Types of aspects extracted by LDA, BTM, and eBTM combined using the Naver dataset.

Aspect	Example Terms
Feeling	Disappointment, Feeling, Goosebumps, Impression, Nervousness, Tears, Tension, Thinking, Thrill
Acting	Acting, Acting ability, Actor, Actor (other term), Gong Yu (actor), Im Siwan (actor)
Direction	Direction, Director, Lee JaeGyu (director), Production, Work (of art), Yoon JongBin (director)
Screenplay	Contents, Dialogue, Lines, Script, Story, Words
Character	Character, General (soldier), Lawyer, Lee SoonShin (character), Main character, Person
Development	Development, Last part, Middle part, Part, Reversal, Starting, Twist
Real life relatedness	Citizen, Country, Father, Friend, Reality, The president, We (Korean culture)

Table 5
Aspects extracted by and aspect extracting power of LDA, BTM, ASUM, and eBTM.

Naver dataset	Terms extracted	Average Terms	Average Unique Terms
LDA	Acting, Actor, Story, Acting, Story, Actor, Acting, Contents, Development, We, Person, Reality, Words, Impression, Tears, Acting, Friend, Director	1.8	**1.3**
BTM	We, Friend, Acting, Acting, Actor, Story, Story, Actor, Director, Acting, Actor, Direction, Friend, Feeling, Acting, Main character, Person, Director, We, Country, Citizen	2.1	1.2
eBTM	Story, Actor, Acting, Direction, Impression, Tears, Acting, Actor, Story, Acting, The president, Actor, Story, We, Thinking, Father, Friend, Acting, Impression	**2.2**	**1.3**
Rotten Tomatoes dataset			
ASUM	Funny, Story, Action, Performance, Love, Story, Act, Performance, Funny, Love, Funny, Funny, Life, Story, Boring, Action, Plot, Time, Time, Man, Rocky	1.05	0.6
eBTM	Damon, Matt, Story, Story, Oyelowo, Deadpool, Cast, Comedy, Coen, Time, Action, Story, Max, Performance, Hanks, Tom, Comedy, Vampire, Fun, Pixar, Story, Star, Wars, Story, Performance	**2.5**	**1.9**

the review summarization technique. Blue Dragon Film Awards is an annual movie awards ceremony that is presented by Sports Chosun in South Korea. We compare the Best Director, Best Actor, and Best Screenplay award winners to movies with screening dates similar to the award winners and with more than ten thousand reviews in Naver movie. Table 7 shows the movies compared and their details. There is a clear difference between the ratings of The Admiral and Kundo, and between the ratings of A Hard Day and The Fatal Encounter. But there is a small difference between the ratings of The Attorney and The Suspect. In this kind of cases, it is also very interesting to see where one movie excels and to differentiate two movies with similar ratings. Hence, we compare movies with similar overall ratings.

We gather ten thousand movie reviews from the six different movies and combine them all into one dataset. This large dataset is then used for the construction of a movie review. Evaluation of the models using this data is shown briefly below. Table 8 shows the evaluation metrics for the sentiment classifier in comparison with the LingPipe classifiers. Table 9 shows the evaluation metrics for the aspect extraction model in comparison with the BTM and LDA. We note that ASUM cannot be included in both evaluations because of the lack of a gold standard list of sentiment seed words. It can be seen that even when using this dataset, the multi-level classifier still outperforms the LingPipe classifiers and eBTM also still outperforms both BTM and LDA.

Using this dataset, review summaries of each of the six movies are constructed using the following method. First, we combine the document-aspect distribution (θ) from the aspect extraction model and the sentiment score of the document from the sentiment classification model, to create the sentiment/aspect result table (as shown in the upper left corner of Fig. 4). Next, we assign names of the aspects using the words in the aspect-word distribution (φ). Using these two tables, we get the documents that are assigned a given aspect and get the average sentiment score of the given aspect. We use two thresholds: one is for the probability assignment of an assigned document in the sentiment/aspect result table (set to 0.9) and another one is for the number of assigned documents for an aspect to be considered in the review summary (set to 30).

Table 10 presents the review summaries of the six movies. There are seven unique aspects extracted, one of them is a weak aspect (direction aspect). Some of the movies do not have a sentiment score on a specific aspect; the number of assigned documents for a specific aspect in a certain movie did not reach the threshold that we set. This is apparent because reviews of a certain movie might not mention topics regarding a certain kind of aspect. For example, both The Suspect and A Hard Day do not have a history aspect because both movies are not in any way based on a past incident.

We discuss the effectiveness of the proposed review summarization framework in two ways. First, we look at the comparisons between movies that won awards and movies that did not win any awards. Next, we compare movies with overall ratings similar to each other to show how movies differ in aspect even if they have similar ratings.

4.3.1. Movies with awards versus movies with no awards

This section shows how the review summarization framework corresponds to the results of Blue Dragon Film Awards 2014. The comparison is shown in Table 10. It is clearly shown that the sentiment score of The Admiral: Roaring Currents for the direction aspect is better than that of Kundo: Age of the Rampant. The same can be seen between A Hard Day and The Fatal Encounter on their sentiment scores for the screenplay aspect. In the case of the acting aspect, both The Attorney and A Hard Day won awards in acting. Both movies' sentiment score on the acting aspect came out higher compared to that of The Suspect and The Fatal Encounter.

4.3.2. Movies with similar ratings

In this section, we show the comparison between four movies with similar higher overall review score: The Admiral: Roaring Currents, The Attorney, The Suspect, and A Hard Day, and two movies with similar lower overall review score: Kundo: Age of the Rampant and The Fatal Encounter. Table 11 shows a modified view of Table 10 to make these comparisons clearer. In the first four movies, we can clearly see that A Hard Day, which garnered the highest overall rating, got the highest sentiment scores on every aspect it extracted. What is interesting in this comparison is that although The Attorney garnered the higher overall rating compared

Table 6
Aspects extracted by and aspects diversity scores of LDA, BTM, and eBTM.

	LDA		BTM		eBTM	
Aspect 1	Again, Just, Really, See, Times	No topic	Thing, That, Really, **We**, Way	Weak	**Story**, **Actor**, **Acting**, **Direction**, Thing	Ambiguous
Aspect 2	**Acting**, **Actor**, Really, **Story**, Very	**Acting**	**Friend**, **Acting**, That, Thing, Really	Ambiguous	Score, Movie score, This, Why, Little bit	No topic
Aspect 3	**Acting**, Best, Super, Very, Really	Weak	**Acting**, **Actor**, **Story**, Really, Very	**Acting**	Forced, **Impression**, **Tears**, Really, Thing	Feeling
Aspect 4	**Story**, **Actor**, **Acting**, **Contents**, **Development**	Ambiguous	**Story**, Point **Actor**, This **Director**	Ambiguous	**Acting**, Best, **Actor**, Really **Story**	**Acting**
Aspect 5	Score, Review score, This, Why, Just	No topic	**Acting**, **Actor**, Point, **Direction**, Forced	**Acting**	Very, See, **Acting**, When, Thing	Weak
Aspect 6	Very, Review score, Why, Thing, Time	No topic	**Friend**, **Feeling**, Thing, Series, **Acting**	Ambiguous	**The president**, Thing, This, **Reality**, **Director**	**Real life relatedness**
Aspect 7	**We**, That, Thing, **Person**, Way	Ambiguous	Really, Why, This, Ah, **Main character**	Weak	**Acting**, **Actor**, **Story**, Very, A little bit	**Acting**
Aspect 8	Why, Very, **Reality**, Little bit, **Words**	Ambiguous	Just, As it is, Thing **Person**, **Director**	Ambiguous	**We**, That, **Thinking**, Way, **Father**	**Real life relatedness**
Aspect 9	**Impression**, Interesting, **Tears**, Really, **Acting**	**Feeling**	**We**, **Country**, This, **Citizen**, That	**Real life relatedness**	Just, **Friend**, Why, A little bit, This	**Weak**
Aspect 10	**Friend**, **Director**, Thing, That, Just	Ambiguous	Movie score, Score, This, Really, Why	No topic	**Acting**, **Impression**, Thing, This, See	Ambiguous
Total		2		3		**5**
Unique total		2		2		**3**

Table 7
Blue Dragon Film Awards 2014 award-winning movies and movies to compare.

Movie	**Award**	**Release date**	**#Movie reviews**	**Overall rating (Naver)**
The Admiral: Roaring Currents	Best Director	July 30, 2014	65,978	8.49
Kundo: Age of the Rampant	–	July 23, 2014	25,804	6.92
The Attorney	Best Actor	December 18, 2013	93,625	8.96
The Suspect	–	December 24, 2013	10,282	8.34
A Hard Day	Best Screenplay; Best Actor	May 29, 2014	13,853	8.99
The Fatal Encounter	–	April 30, 2014	17,040	7.05

Table 8
Sentiment classification model evaluation using the Blue Dragon dataset.

	LingPipe (UNI)	**LingPipe (BI)**	**Our own**
Precision	89.3	93.9	**99.0**
Recall	**90.1**	88.9	85.8
F-Score	89.7	91.4	**92.0**
Accuracy	82.8	85.1	**85.5**

Table 9
Aspect extraction model evaluation using the Blue Dragon dataset.

	Aspect diversity scores		Aspect extracting power scores	
Models	Number of aspects	Number of unique aspects	Average number of aspect terms	Average number of unique aspect terms
LDA	5	5	2.1	1.9
BTM	6	3	2.4	1.9
eBTM	7	6	2.4	2.1

Table 10
Review summaries of the six movies.

Movie	Screenplay	History	Feeling	Direction	Patriotism	Character	Acting
The Admiral: Roaring Currents	86.45	88.21	92.42	**73.25**	95.06	77.83	94.93
Kundo: Age of the Rampant	66.02	71.03	68.71	58.75	—	58.61	80.41
The Attorney	—	80.46	89.37	–	92.97	73.96	**95.16**
The Suspect	71.53	—	82.34	62.77	—	65.08	94.59
A Hard Day	**91.03**	—	93.92	—	—	90.47	98.47
The Fatal Encounter	62.99	76.69	70.40	49.05	84.11	—	85.01

Table 11
Modified view of Table 10.

Movie	Overall rating	Screenplay	History	Feeling	Direction	Patriotism	Character	Acting
A Hard Day	8.99	**91.03**	—	**93.92**	—	—	**90.47**	**98.47**
The Attorney	8.96	—	80.46	89.37	—	92.97	73.96	95.16
The Admiral: Roaring Currents	8.49	86.45	**88.21**	92.42	73.25	**95.06**	77.83	94.93
The Suspect	8.34	71.53	—	82.34	62.77	—	65.08	94.59
The Fatal Encounter	7.05	62.99	**76.69**	**70.40**	49.05	84.11	—	**85.01**
Kundo: Age of the Rampant	6.92	**66.02**	71.03	68.71	**58.75**	—	58.61	80.41

to The Admiral: Roaring Currents, The Admiral: Roaring Currents got higher sentiment scores compared to The Attorney's sentiment scores, with the exception on the scores on the acting aspect. We can infer that The Attorney's high overall rating is attributed to the performance of the actors in the movie.

Another possible comparison is the comparison between two lower-rated movies, The Fatal Encounter and Kundo: Age of the Rampant. The movies' ratings are quite similar to each other, but through the review summarization, we can see the difference between two movies. The Fatal Encounter got higher sentiment scores on the history, feeling, and acting aspect, while Kundo: Age of the Rampant got higher sentiment scores on screenplay and direction aspect.

5. Conclusion

The present study proposed an adaptable fine-grained sentiment extraction algorithm for short texts. We demonstrated that the combination of both natural language processing and machine learning techniques are more effective than a single-level classifier. We also presented that neural network algorithms are better suited for sentiment classification in terms of precision, recall, and F-measure. The resulting sentiment classifier outperforms existing approaches on several datasets collected from multiple domains and in different languages, thus proving the model's adaptability.

We also extend the biterm topic model for short texts into eBTM. We report that eBTM outperforms LDA and BTM in terms of aspect diversity and outperforms LDA, BTM and ASUM in terms of aspect extracting power. The experimental results indicate that the proposed technique performs well on short texts where ASUM performs poorly on short texts. Our model is not dependent on external seed words and it is usable on different domains and languages. These results support our hypothesis that it is more appropriate to separate aspect extraction and sentiment analysis for adaptability.

Finally, we show an effective application of our fine-grained sentiment extraction model to summarization of multiple online reviews. Using the results of a film award ceremony, we show the effectiveness of our online review summarization framework in two ways. We present that movies that received an award have a higher aspect sentiment score compared to those without an award. We also demonstrate that our review summarization framework can differentiate movies with similar overall sentiment rating.

Two limitations of the study are as follows. First, although the proposed method is more adaptable than previous methods, it is not very adaptable to languages that do not have support on tokenizers and POS taggers. In this paper, tokenizers and POS taggers are used extensively to classify sentiments and extract aspects. However, we posit that building a simple tokenizer and POS tagger would suffice. Another limitation is that we need to manually tag the aspects with aspect labels in order to view the final review summary. This is considered to be another challenge in topic models [10] called topic labeling problem. Since in this paper we used an extension of LDA to extract aspects, the proposed method is also bound to the said limitation. However, there are proposed solutions to this problem [41] which are available to use and easily extensible to our method.

As follow-up studies, we plan to improve eBTM in several ways. We can include a background distribution to remove noise from short texts. This noise includes stopwords and common words from a specific domain that do not have importance in the extraction of aspects (i.e. “movie” and “product”). Another way is to integrate ASUM's idea on sentiment seed words [6] and find out whether it improves ASUM's performance on short texts or not. However, this contradicts to the adaptability of eBTM, thus we also need to find a way to automate the construction of a sentiment seed word list. Finally, we plan on applying this framework to analyze Twitter data and its capability to extract aspects and their sentiments and use them for review summarization. Using Twitter, we can do a lot of wide and global analysis with regards to the products and reviews and their aspects. One example of this application is to compare summaries of reviews of a single product that is sold in multiple countries using Twitter data. Since Twitter can distinguish tweets from different countries, this is a possible application of the review summarization framework.

Notes

1. https://github.com/twitter/twitter-korean-text
2. http://movie.naver.com/movie/bi/fi/prize.nhn?code=18 & rnd=35

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046711)

References

[1] I.E. Vermeulen, D. Seegers, Tried and tested: the impact of online hotel reviews on consumer consideration, Tourism Manag. 30 (1) (2009) 123–127.
[2] Q. Ye, R. Law, B. Gu, The impact of online user reviews on hotel room sales, Int. J. Hosp. Manag. 28 (1) (2009) 180–182.
[3] W. Duan, B. Gu, A.B. Whinston, Do online reviews matter??An empirical investigation of panel data, Decis. Support Syst. 45 (4) (2008) 1007–1016.
[4] P. Chatterjee, Online reviews: do consumers use them?, Adv. Consum. Res. 28 (2001) 1.
[5] I. Titov, R.T. McDonald, A joint model of text and aspect ratings for sentiment summarization, In ACL 8 (2008) 308–316.
[6] Jo Y, Oh AH. Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining 2011 Feb 9 (pp. 815-824). ACM.
[7] S. Kim, J. Zhang, Z. Chen, A. Oh, S. Liu, A Hierarchical Aspect-Sentiment Model for Online Reviews. In AAAIJul 16, 2013.
[8] A. Bagheri, M. Saraee, F. De Jong, ADM-LDA: an aspect detection model based on topic modelling using the structure of review sentences, J. Inf. Sci. 40 (5) (2014) 621–636.
[9] J. Read, Using emoticons to reduce dependency in machine learning techniques for sentiment classification (In Proceedings of the ACL student research workshop), Association for Computational Linguistics, 2005, pp. 43–48.
[10] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[11] X. Yan, J. Guo, Y. Lan, X. Cheng, A biterm topic model for short texts. in: Proceedings of the 22nd international conference on World Wide WebMay 13 pp. 1445–1456. International World Wide Web Conferences Steering Committee, 2013.
[12] N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin, A.F. Smeaton, Topic-dependent sentiment analysis of financial blogs (In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion), ACM, 2009, pp. 9–16.
[13] R. Dehkharghani, B. Yanikoglu, D. Tapucu, Y. Saygin, Adaptation and use of subjectivity lexicons for domain dependent sentiment classification (In Data Mining Workshops (ICDMW)(, 2012 IEEE 12th International Conference on), IEEE, 2012, pp. 669–673.
[14] H. Kanayama, T. Nasukawa, Fully automatic lexicon expansion for domain-oriented sentiment analysis (In Proceedings of the 2006 Conference on Empirical

Methods in Natural Language Processing), Association for Computational Linguistics, 2006, pp. 355–363.
[15] J. Read, J. Carroll, Weakly supervised techniques for domain-independent sentiment classification (In Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion), ACM, 2009, pp. 45–52.
[16] V. Raychev, P. Nakov, Language-Independent Sentiment Analysis Using Subjectivity and Positional Information. In RANLPSep 14 pp. 360–364, 2009.
[17] A.L. Gînscă, E. Boroş, A. Iftene, D. TrandabAˇţ, M. Toader, M. Corîci, C.A. Perez, D. Cristea, Sentimatrix: multilingual sentiment analysis service (In Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis), Association for Computational Linguistics, 2011, pp. 189–195.
[18] H. Tanev, B. Pouliquen, V. Zavarella, R. Steinberger, Automatic expansion of a social network using sentiment analysis (In Data Mining for Social Network Data), Springer US, 2010, pp. 9–29.
[19] A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision, CS224N Project Report, Stanford. 1 (2009) 12.
[20] E. Kouloumpis, T. Wilson, J.D. Moore, Twitter sentiment analysis: the good the bad and the omg!, ICWSM 11 (2011) 538–541.
[21] D.H. Yang, G. Yu, A method of feature selection and sentiment similarity for Chinese micro-blogs, J. Inf. Sci. (2013) (Mar 18:0165551513480308).
[22] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, A. Kappas, Sentiment strength detection in short informal text, J. Am. Soc. Inf. Sci. Technol. 61 (12) (2010) 2544–2558.
[23] M.E. Basiri, N. Ghasem-Aghaee, A.R. Naghsh-Nilchi, Exploiting reviewers' comment histories for sentiment analysis, J. Inf. Sci. 40 (3) (2014) 313–328.
[24] Y. Bae, H. Lee, Sentiment analysis of Twitter audiences: measuring the positive or negative influence of popular twitterers, J. Am. Soc. Inf. Sci. Technol. 63 (12) (2012) 2521–2535.
[25] H.K. Aldayel, A.M. Azmi, Arabic tweets sentiment analysis–a hybrid scheme, J. Inf. Sci. (2015) (Oct 19:0165551515610513).
[26] Balahur A. Sentiment analysis in social media texts. In4th workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis 2013 Jun 14 (pp. 120-128).
[27] M. Thelwall, K. Buckley, Topic-based sentiment analysis for the social web: the role of mood and issue-related words, J. Am. Soc. Inf. Sci. Technol. 64 (8) (2013) 1608–1617.
[28] S.M. Jiménez-Zafra, M.T. Martín-Valdivia, E. Martínez-Cámara, L.A. Ureña-López, Combining resources to improve unsupervised sentiment analysis at aspect-level, J. Inf. Sci. (2015) (Jul 9:0165551515593686).
[29] Q. Miao, Q. Li, D. Zeng, Fine-grained opinion mining by integrating multiple review sources, J. Am. Soc. Inf. Sci. Technol. 61 (11) (2010) 2288–2299.
[30] T.T. Thet, J.C. Na, C.S. Khoo, Aspect-based sentiment analysis of movie reviews on discussion boards, J. Inf. Sci. (2010) (Nov 15:0165551510388123).
[31] A. Karandikar Clustering short status messages: A topic model based approach (Doctoral dissertation, University of Maryland).
[32] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, R. Van de Walle, Using topic models for twitter hashtag recommendation. in: Proceedings of the 22nd international conference on World Wide Web companionMay 13 pp. 593–596. International World Wide Web Conferences Steering Committee, 2013.
[33] E.H. Kim, Y.K. Jeong, Y. Kim, K.Y. Kang, M. Song, Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news, J. Inf. Sci. (2015) (Oct 5:0165551515608733).
[34] Y. Zhu, L. Li, L. Luo, Learning to classify short text with topic model and external knowledge (In Knowledge Science, Engineering and Management), Springer Berlin Heidelberg, 2013, pp. 493–503.
[35] L. Hong, B.D. Davison, Empirical study of topic modeling in twitter (In Proceedings of the first workshop on social media analytics), ACM, 2010, pp. 80–88.
[36] Xu Z, Lu R, Xiang L, Yang Q. Discovering user interest on twitter with a modified author-topic model. In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on 2011 Aug 22 (Vol. 1, pp. 422-429). IEEE.
[37] T. Lin, W. Tian, Q. Mei, H. Cheng, The dual-sparse topic model: mining focused topics and focused terms in short text (In Proceedings of the 23rd international conference on World wide web), ACM, 2014, pp. 539–550.
[38] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, D. McClosky, The Stanford CoreNLP Natural Language Processing Toolkit. In ACL (System Demonstrations)Jun 23 pp. 55–60, 2014.
[39] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin, LIBLINEAR: a library for large linear classification, J. Mach. Learn. Res. 9 (2008) 1871–1874.
[40] B. Baldwin, B. Carpenter, LingPipe. Available from World Wide Web: 〈http://alias-i.com/lingpipe〉, 2003.
[41] Mei, Q., Shen, X. and Zhai, C., 2007, August. Automatic labeling of multinomial topic models. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 490-499). ACM.