신뢰성있는 온라인 고객 리뷰 텍스트 마이닝 기반 식당 개별 음식 아이템 평가

무자밀 후세인 사이드*, 정선태**

*숭실대학교 정보통신공학과, **숭실대학교 스마트시스템소프트웨어학과

e-mail: engr.muzamilshah@gmail.com, cst@ssu.ac.kr

Rating Individual Food Items of Restaurant Menu based on Online Customer Reviews using Text Mining Technique

Muzamil Hussain Syed*, Sun-Tae Chung**

* Dept. of Information and Telecommunication, Graduate School, Soongsil University.

**Dept. of Smart System Software, Soongsil University

Abstraction

The growth in social media, blogs and restaurant listing directories have led to increasing customer reviews about restaurants, their quality of food items and services available on the internet. These user reviews offer a massive amount of valuable information that can be used for various decision-making purposes. Currently, most food recommendation sites provide recommendation scores about restaurants rather than food items of the restaurant and the provided recommendation scores may be biased since they are calculated only from user reviews listed only in their sites. Usually, people wants a reliable recommendation about foods, not restaurant. In this paper, we present a reliable Korean food items rating method; we first extract food items by applying NER technique to restaurant reviews collected from many Korean restaurant recommendation web sites, blogs and web data. Then, we apply lexicon-based sentiment analysis on collected user reviews and predict people's opinions as sentiment polarity scores (+1 for positive; -1 for negative; 0 for neutral). Finally, by taking average of all calculated polarity scores about a food item, we obtain a rating to individual menu items of the restaurant. The proposed food item rating is more reliable since it does not depend on reviews of only one site.

1. Introduction

Sentiment Analysis or opinion mining from texts can be seen as a natural language processing (NLP) task that aims to analyze opinions, sentiments, and emotions expressed in unstructured data [1]. A common task in this research area is polarity classification, which consists in classifying the overall sentiment present in a document or sentence. Usually this task is simplified by classifying a text or a sentence in 3 classes: positive, negative or neutral. To build a sentiment classifiers, two main approaches have been investigated: lexicon-based methods [3], and machine learning algorithms [4,5].

Applying sentiment analysis to user reviews, to know their opinion about entities can give us some useful insights for marketing and decision-making purposes.

Analyzing and extracting meaningful information from the user reviews is useful from both client and restaurant owner perspectives. Today, people are highly interested in searching client's reviews on restaurants online to know one's perception of the food quality and services of a restaurant before the visit. The restaurant owner also gets to know about the clients' opinion over the quality of food items and services offered. This can help the restaurant owner to improve its marketing strategy and the quality of food and services. Most of current food recommendation sites in Korea provide recommendation scores about restaurants rather than food items of the restaurant, furthermore the provided recommendation scores may not be reliable since they are calculated from user views listed only in their sites.

In this paper, we present a reliable food items rating method, which is based on text analysis on user reviews from different sources; food recommendation web sites, blogs, SNS(twitter, instagram, etc). Thus, it can provide a way to the local restaurant owner to identify people's opinions about the quality of their food and help other users to find the best food place for dining in. We use the Named Entity Recognition (NER) technique to identify food item names in restaurant reviews and apply lexicon-based sentiment analysis to extract people's opinions by sentiment polarity. In our work, we utilize a lexical resource approach such as a dictionary of opinionated terms. Korean Sentiment Analysis Corpus named KOSAC [7] is one such resource, which is the sentimental dictionary for Korean words and assigns three sentiment numerical scores to each word as positivity, negativity and neutral.

2. Proposed Method

The proposed method consist of three main steps.

2.1 Data Collection

The system initially targets for a small business area near Isu station in Seoul. We collects local restaurant data and reviews from multiple sources which are broadly categorized into two categories. i) Concrete data source: The data source contains complete information about the restaurant, menus, and reviews. These are GooglePlaces, KakaoPlace, MangoPlate, Naver Store, SiksinHot and DiningCode. ii) Partial data source: The data source which provides user opinion data by their posts, blogs or comments about the food items of the restaurant. These are Twitter, Instagram, Naver blogs.

The Collected data set includes 32 restaurants and 16656 reviews. The collected data is stored in the MongoDB database. We creat a list of 20 famous selected Korean food items that contain only the real common food item names, that are used to identify the food name from the review and used as Named Entity Recognition. This is important as many of the restaurant menu names are customized based on the recipe, but users while writing a review, uses the common name of the food items. Figure1 shows the architecture of data collection and storage.

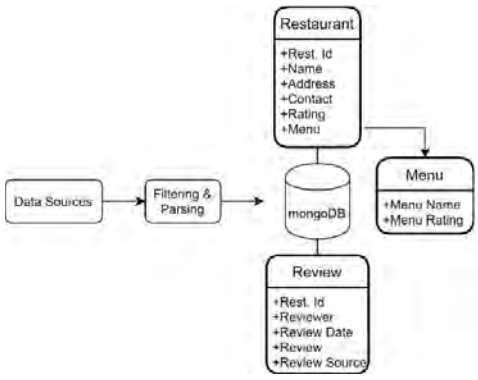

Figure 1: Data collection and storage architecture.

2.2 Rating Food Item

The initial rating of food items in the restaurant menu is applied based on the overall average rating score of the restaurant. The average rating of the restaurant is calculated from the rating scores obtained from the concrete data source. For example, if the rating scores for a restaurant *R* obtained from different data sources is 4.1, 2.6, 4.2, 4.3 and 3.6, then the overall average rating would be 3.8 , which is assigned as the initial rating score for individual food items in the restaurant menu. This rating is then further updated based on the opinion of the user reviews. The process of rating food items in restaurant menu list consists of a pipeline of operations; see Figure 2.

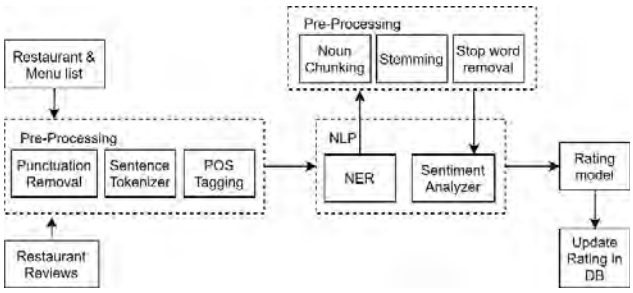

Figure 2: Process of rating food item.

The restaurant's menu list and reviews are fetched from the database.

A. Pre-processing of ReviewData

The aim of preprocessing is to remove unwanted and noisy data. To pre-process the review data, Korean NLP in Python KoNLPy has been used. KoNLPy is a Python package for natural language processing (NLP) of the Korean language. The pre-processing steps are performed at two stages. The need for processing data into two stages arises from the requirement of extracting the food names from the reviews. By processing the data in one stage may results in non identified food names from the restaurant menu. The first stage comprises the following tasks:

i. *Punctuation Removal*: Punctuations in a text generally do not provide any useful information. This step, therefore, erases the punctuation characters from the word.

ii. *Tokenization & POS tagging:* This process breaks a stream of text into a list of words. Each review is broken into sentences. The sentence detector in the KoNLPy toolkit is used to break reviews into sentences. Then, each of these sentences is tokenized and POS tagged using the toolkit.

iii. *Noun Chunking:* The POS annotated tokens are then sent to the chunker, which combines the noun token and other possible noun category token identified by the toolkit. The noun chunker uses Regex to combine the noun token.

```
Regex = """
NP: {<N.*>*<UN>?<N.*>?<Suffix>?}
 """
```

The Regex considers all the noun tags (보통명사, 고유명사, 일반 의존 명사, 단위 의존 명사) and other tags that are estimated to be a noun (명사추정범주). This is because the food names in the reviews are not just the standard food names included in the food menu

The second stage of data pre-processing is performed after NER extraction to perform sentiment analysis. The second stage comprises the following tasks:

iv. *Stemming*: This step reduces words to its stem or root form as stemming simplifies the sentiment analysis process. The same word can be used in a different flavor for grammatical reasons such as consult, consulting, consultant.

v. *Stop Word Removal*: Stop words consist of prepositions, help verbs, articles, and so forth. They typically do not contribute to analyzing sentiments and are removed from the text.

B. Named Entity Recognition (NER)

Named Entity Recognition NER also known as entity chunking/extraction, is a popular technique used in information extraction to identify and segment the named entities, and classify or categorize them under various predefined classes. After pre-processing, an NER technique is utilized to extract food names from customer reviews. The noun chunks don't fit enough to find out food names from the reviews. To make use of NER to identify food names in restaurant reviews, a corpus annotated with food names is

required. Since such corpus was not available in Korean language for food names, we had to create one. Different approaches could be applied to make a custom food name annotated corpus[2] for Korean language. The possible approach could be, annotating all food names in collected reviews and train the model using spaCy to detect food domain name automatically. Instead of creating an annotated corpus, we use a list of selected food items that could be used as a NER. To identify a food name entity, we have used a two-way matching approach.

Firstly, in the selected list of food items only the common names are defined; i.e. '초밥', '치킨', '갈비' etc. The selected food item list is used to detect the noun chunks that contain the food word from the review texts. In most of the cases the food names in restaurant menus are more verbose and customized based on the recipe but users, while writing a review, use the common name of the food items; i.e '초밥' that makes it harder to identify the food item from menu list; i.e '스페셜 초밥', '로로초밥', '특 초밥'.

After a noun chunk matched with a food name from the selected list, the noun chunk is then compared with the food item names in the restaurant menu list using a fuzzy string matching algorithm to find the Levenshtein similarity ratio based on the Levenshtein distance. The similarity measure recognizes the menu items accurately and ignores the problem of spaces and wrong word tag collected by noun chunk.

Menu item	Noun Chunk	Similarity Ratio	Description
스페셜 초밥	과 스페셜 초밥	86 %	Extra noun token
특 초밥	특초밥	86 %	Space missing
로로초밥	로초밥	86%	Token missing
회덮밥	회덮밥	100%	Full matched

Table 1: Similarity measure of menu items collected from review.

The sentences containing the food item names are collected and further processed for sentiment analysis.

C. Sentiment Analysis

Sentiment Analysis is the computational methodology used in the study of sentiments or opinions of people towards various entities like individuals, subjects or events [1]. Suppose, we have a product review, it figures out if the review is of positive polarity or negative polarity.

Many different techniques have been presented for analysis of sentiments in product reviews. These approaches are basically categorized into machine learning based and lexicon-based approaches. Machine learning based approaches include some supervised and unsupervised classification algorithms. Lexicon based methodologies consist of dictionary-based and corpus-based approaches.

In our work we have applied a lexicon-based approach in order to avoid the need to prepare a labeled training set. The main disadvantage of machine learning approach is their reliance on labeled data. It is extremely difficult to ensure that sufficient and correctly labelled data can be obtained. In this work, we have used the subjectivity lexicon MPQA corpus named KOSAC [7]. A subjectivity lexicon is a list of positive or negative opinion words.

Korean Sentiment Analysis Corpus (KOSAC) consists of 332 news articles taken from the Sejong Syntactic Parsed Corpus [6]. The corpus includes 7,713 sentence subjectivity tags and 17,615 opinionated expression tags based on the annotation scheme called KSML which reflects the characteristics of the Korean language. Examples of sentiment scores associated with KOSAC entries are shown in Figure 3.

ngram	freq	COMP	NEG	NEUT	POS	max.value	max.prop
가/JKS	1	0	0	0	1	POS	1
가/JKS; 있/VV	1	0	0	0	1	POS	1
가/JKS; 있/VV; 었/EP	1	0	0	0	1	POS	1
가/VV	3	0	0	0	1	POS	1
가/VV; ㄴ다 /EF	1	0	0	0	1	POS	1

Figure 3: KOSAC tagging.

In our approach, we apply sentence-level scoring. Therefore, it is important that a phrase must contain a single food item. If a sentence contains multiple food items names, we first split the sentence into separate sentences to have only a single food item.

The sentiment analyzing algorithm takes POS tagged sentence as input and provides a scoring value for the positive, negative and neutral polarity of the sentence. The overall sentiment of the sentence is calculated based on the higher value of polarity. For example, if a positive polarity score is greater, the overall sentiment of the sentence is 1, for negative and neutral polarity would be -1 and 0 respectively. The overall sentiment score is then applied to the respective food item in the sentence.

The overall rating of the food menu item from the reviews is then calculated based on the obtained sentiment score using the following formula.

$$Rating\ from\ reviews = \sum_{k=1}^{n} \frac{(sentiment\ score * 5)}{food\_item_{count}}$$

The calculated food item rating from reviews then applied to initial rating of the restaurant menu item to get final rating of the food item.

$$FinalRating_{food\_item} = \frac{(initial\_rating + review\_rating)}{2}$$

3. Results

The results from a restaurant reviews are shown in Figure 4, and 5. The sentences are taken from the reviews which contain those food item names which we are interested to evaluate and rating those food items in the restaurant menu list.

The words polarity shown in Figure 4, depicts % in terms of positive, negative and neutral words. However, positive words holds high ratio than the others & neutral remains with the least.

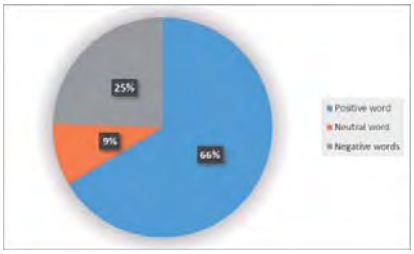

Figure 4: Words polarity.

The overall total rating is shown in Figure 5, the calculation of the total rating score is made on the basis of its initial rating and the calculated rating from the user reviews.

Figure 5: Initial, review and overall rating of food items.

The word cloud shown in Figure 6, visually shows the frequent words in review's sentences which contain a food item name.

Figure 6: Word cloud for frequent words in the sentences.

4. Coclusion and Future Work

Sentiment analysis is the process of identifying the feeling expressed in the text or document. We proposed a system for mining the restaurant and reviews data and to calculate score for the food items in the restaurant menu. It is evaluated the the proposed system has produced reasonable results in calculating the ratings for food items from user reviews. In future work, the NER technique would be applied on trained annotated corpus to identify all kinds of food items with more accuracy. A hybrid approach would be adopted by combining machine learing approach with existing approach to extract more features to handle implicit sentiment analysis and to identify deceptive reviews.

References

[1] Lei Zhang and Bing Liu: "Sentiment Analysis and Opinion Mining" Encyclopedia of Machine Learning and Data Mining 2017.

[2] Burusothman A., Paraneetharan S., Sanjith B., Thamayanthy S., Surangika R.: "Ruchi: Rating Individual Food Items in Restaurant Reviews" Intl. Conference on Natural Language Processing 2015.

[3] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: "Lexicon-based methods for sentiment analysis." Computational linguistics 2011.

[4] Pang B., Lee L., Vaithyanathan S.: "Thumbs up? Sentiment classification using machine learning techniques. Association for Computational Linguistics 2002.

[5] Suchita V., Sachin N., Deshmukh: "Sentiment classification using machine learning techniques. International Journal of Science and Research 2013.

[6] Hayeon J., Munhyong K., Hyopil S.: "KOSAC: A Full-Fledged Korean Sentiment Analysis Corpus" PACLIC 2013.

[7] http://word.snu.ac.kr/kosac/lexicon.php