



저작자표시-비영리-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#) 

공학석사 학위논문

Sentiment Polarity Classification of Comments on Korean News Articles Using Feature Reweighting

가중치 조정을 이용한

한국어 신문 기사의 댓글에 대한 감정 이진 분류



지도교수 김재훈

1945

2009년 8월

한국해양대학교 대학원

컴퓨터공학과

서형원

본 논문을 서형원의 공학석사 학위논문으로 인준함

위원장 공학박사 류길수 인

위원 공학박사 박휴찬 인

위원 공학박사 김재훈 인



2009년 8월

한국해양대학교 대학원

컴퓨터공학과

서형원

Contents

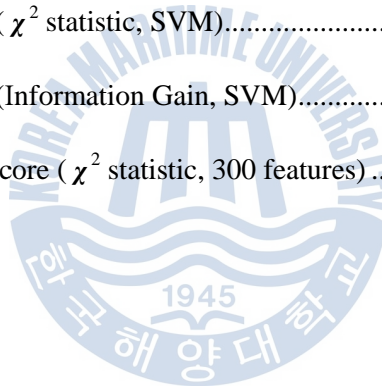
List of Figures	iii
List of Tables	iv
Abstract	v
Chapter 1 Introduction	1
Chapter 2 Related Works	4
2.1 Sentiment Classification.....	4
2.2 Feature Weighting in Vector Space Model	5
2.3 Feature Extraction and Selection.....	7
2.4 Classifiers	10
2.5 Accuracy Measures	14
Chapter 3 Feature Reweighting	16
3.1 Feature extraction in Korean	16
3.2 Feature Reweighting Methods.....	17
3.3 Examples of Feature Reweighting Methods.....	18
Chapter 4 Sentiment Polarity Classification System	21
4.1 Model Generation.....	21
4.2 Sentiment Polarity Classification	23
Chapter 5 Data Preparation	25
5.1 Korean Sentiment Corpus	25

5.2 Korean Sentiment Lexicon.....	27
Chapter 6 Experiments	29
6.1 Experimental Environment.....	29
6.2 Experimental Results.....	30
Chapter 7 Conclusions and Future Works	38
Bibliography	40
Acknowledgments	45



List of Figures

Figure 3.1: An example of feature reweighting for α and β	19
Figure 4.1: System architecture for the model generation.....	22
Figure 4.2: System architecture for the sentiment polarity classification.....	24
Figure 6.1: Performance according to classifiers	32
Figure 6.2: Overall F_1 score (Document Frequency, SVM).....	33
Figure 6.3: Overall F_1 score (χ^2 statistic, SVM).....	34
Figure 6.4: Overall F_1 score (Information Gain, SVM).....	35
Figure 6.5: The result of F_1 score (χ^2 statistic, 300 features).....	36



List of Tables

Table 2.1: Illustration of the classification context table.....	14
Table 5.1: The statistics on constructed corpus.....	26
Table 5.2: Categories of documents.....	27
Table 5.3: The number of total sentiment words.....	28



가중치 조정을 이용한
한국어 신문 기사의 댓글에 대한 감정 이진 분류

서 형 원

한국해양대학교 대학원

컴퓨터공학과

지도교수 김 재 훈



초록

일반적으로 인터넷 신문 기사에 대한 댓글은 그 신문 기사에 대한 주관적인 감정이나 의견을 포함하고 있다. 따라서 이런 신문의 댓글에 대한 감정을 인식하고 분류하는 데에는 그 신문의 원문 내용이 중요한 영향을 미친다. 이런 점에 착안하여 본 논문은 기사의 원문 내용과 감정 사전을 이용하는 가중치 조정 방법을 제안하고, 제안된 가중치 조정 방법을 이용해서 한국어 신문의 댓글에 대한 감정 이진 분류 방법을 제안한다.

가중치 조정 방법에는 다양한 자질 집합이 사용되는데 그것은 댓글에 포함된 감정 단어, 그리고 감정 사전과 뉴스 기사의 본문에 관련된 자질들, 마지막으로 뉴스 기사의 카테고리 정보가 포함되어 있다. 여기서 말하는 감정 사전은 한국어 감정 사전을 의미하며 아직 공개된 것이 없기 때문에, 기존에 있는 영어 감정 사전을 이용하여 구축하였다.

본 논문에서 제안된 감정 이진 분류는 기계 학습을 이용한다. 일반적으로 기계 학습을 위해서는 학습 말뭉치가 필요한데 특별히 감정 분류 문제에서는 긍정 혹은 부정 감정 태그가 부착된 말뭉치가 필요하다. 이 말뭉치의 경우도, 공개된 한국어 감정 말뭉치가 아직 없기 때문에 말뭉치를 직접 구축하였다. 사용된 기계 학습 방법으로는 Naïve Bayes, k -NN, SVM 이 있고, 자질 선택 방법으로는 Document Frequency, χ^2 statistic, Information Gain 이 있다.

그 결과, 댓글 안에 포함된 감정 단어와 그 댓글에 대한 기사 본문이 감정 분류에 매우 효과적인 자질임을 확인할 수 있었다.



Chapter 1

Introduction

Recently, interests on the Internet has rapidly growing up and communication on the Web also has been explosively increasing. For these reasons, interests for automatically mining lots of emotions, opinions, judgments and even recommendations have been increasing too (Liu, 2006). Here are some general questions on these interests: “What is the general opinion about products such as cameras?” and “Which aspects of our products are complained and why?” All these questions can be good reasons why opinion mining and sentiment analysis that deal with the linguistic (or computational) treatment of opinion, sentiment, and subjectivity in documents are needed.

The sentence, for example, *“I love this place! Been loving it for over two years now.”* expresses a positive sentiment. Another sentence *“I honestly don’t get how you can say that.”* expresses a negative sentiment. Both sentences contain subjective opinions. However, the sentence *“The 30GB White iPod (video) is one of Apple’s new (5th) generation of iPod digital media players, featuring video and audio playback, a 30GB hard drive, and a terrific 2.5" diagonal 320 X 240 QVGA color LCD display.”* represents an objective fact. These sentences can be easily found in various forms such as news articles, private blogs, forums, discussion groups or review sites.

Most of those documents can be divided into objective or subjective forms (i.e. they

both of them, so classifying them is difficult and needs more linguistic processing. Nevertheless, sentiment classification is useful for the Web pages that don't have any explicit rating indicators, because opinions of other people are very useful when people want to make a decision for purchasing products or services. Moreover it can be used for filtering out e-mail messages with impolite or abusive words, e.g. it can label a sentiment category to your emails according to whether they express angry or happy emotions (Spertus, 1997). For these reasons, thus, mining or analyzing tons of sentiment expressions automatically in the web pages are very important.

Many researchers have worked various areas of sentiment analysis at the sentence and the phrase level (Wiebe 1999; Wilson 2005), also at the document level (Pang *et al.* 2002; Turney 2002). Several researchers worked about the method that automatically identifying adjectives, verbs, and *n*-grams that associated with sentiment expressions (Turney 2002; Hatzivassiloglou and McKeown 1997; Wiebe 2002; Wiebe *et al.* 2001). And several researchers extracted sentiment expressions using a bootstrapping pattern learning system, also extracted patterns for subjective expressions (Rillof and Wiebe 2003; Pang and Lee 2004). Most these researchers use machine learning algorithms which take feature vectors as inputs and produce a sentiment class like positive or negative. The features are extracted from words in documents and are extended from the words into.

This thesis treats news articles that mainly have comments to classify their sentiment class such as positive or negative, and it assumes that the comments are closely related with body texts in news articles because writers of comments draw them up after reading the body texts. Especially comments of the news articles concerned with politics account for a large proportion of negative sentiments. Thus the features in this thesis contain ones

related with body texts such as words and categories. Based on this assumption, this thesis presents a method for sentiment polarity classification of comments in Korean news articles using machine learning algorithms.

To do this, this thesis builds training data and a Korean sentiment lexicon because they are not available yet in Korean. The training data, namely training corpus, consists of pairs of comments and their corresponding polarities such as positive and negative. The resultant data consists of 1,377 articles, which have 8,320 comments. The Korean sentiment lexicon is made from the English sentiment lexicon using an English-Korean dictionary. To evaluate the proposed method, several classifiers such as Naïve Bayes, k -NN, SVM are used with three feature selection methods such as Document Frequency, χ^2 statistic, Information Gain. The experiments have shown that the performance in case using χ^2 score with SVM is the best. Furthermore this thesis has demonstrated that features related with sentiment words and body texts are effective for sentiment polarity classification of comments in news articles.

The remainder of this thesis is structured as follows. Chapter 2 discusses several related works in brief. Chapter 3 represents a few feature extraction problems and solutions in Korean and a novel feature reweighting method, and then Chapter 4 describes sentiment polarity classification system. Chapter 5 expresses data sets preparation for evaluation. The last chapter in this thesis concludes and discusses results of evaluation.

Chapter 2

Related Works

Sentiment classification aims to determine the opinion of a speaker or a writer with respect to some topics or products. It has emerged as a current research area, but it is still in its introductory stage.

This chapter focuses the related works only on feature weighting in sentiment classification. At first, sentiment classification will be introduced, and then feature vector representation of documents and feature extraction/selection method will be represented. Next, three classifiers, i.e. Naïve Bayes, k -Nearest Neighbor, and Support Vector Machine, will be presented in brief and some evaluation measures will be discussed.

2.1 Sentiment Classification

In a broad sense, as it is mentioned previously, opinion mining means that finding out what the author's private opinion or feeling about the object in a web document is (Pang and Lee, 2008). This object can be a product, a movie, a service, etc. Also it is a task to obtain the overall sentiment properties of a document and to discover details that people like/dislike at the sentence. In fact, this feature-based opinion mining has a lot of useful applications (Liu 2006). For instance, potential customers try to purchase products or services tend to focus on the public opinion. Besides, finding other's opinions from web pages is easier than before. People can easily post reviews of products or services at

private web blogs, internet forums, and threads in review sites. Because these opinions could be exposed easily and have a great influence to people, it is one of a primary factor for opinion mining. As in a similar case, businesses always want to know what customers' opinions about their products or services are. They can collect them automatically and use as feedback. That is, positive opinions or sentiment expressions in a comment might be placed an advertisement of the product.

Sentiment classification, in a little bit different case, treats opinion mining as a text classification problem (Nigam and Hurst 2004). It means this task goes to a document level and classifies a document whether it contains a positive or negative sentiment totally. In general, text classification uses content words based on defined classes and these words are principally a noun. This task classifies documents through the subject such as sports, education, politics, etc. On the other hand, sentiment classification treats sentiment words that consist of adverb or adjective such as excellent, good, bad or poor (Pang *et al.* 2002). In the sentiment classification, these terms, mainly adjectives and adverbs and fixed expressions (e.g. “*dreams come true*”, “*back off*”), are used as sentiment indicators (Rimon 2005). The list of sentiment indicators can be made manually, built semi-automatically using sources such as WordNet (Miller *et al.* 1993), or obtained by machine learning methods from tagged samples in the domain of interest. Finally, using these sets of sentiment indicators helps sentiment classification to classify the document.

2.2 Feature Weighting in Vector Space Model

A vector space model (Salton et al. 1975) is used for ranking documents. Simply, the model represents documents as a vector. Each term in the document is represented to a

dimension of the vector. If a term occurs in the document, its value in vector is non-zero. There are various modifications to compute these values, and several alternatives provide better results than the other approaches (which are based on probability theory).

TF-IDF (Term Frequency-Inverse Document Frequency) weighting (Salton *et al.* 1975) is one of the best known statistical measures and is used when it should be found how important a word is at a document. TF (term frequency) denotes how many times a term has appeared at a document and is derived as follows:

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}, \quad (2.1)$$

where n_{ij} is the number of times that a particular word w_i appears in a document d_j and $\sum_k n_{kj}$ is the total of times that the words in the same document d_j . So the more word is appeared, the more important word might be. However if the word is appeared at the same time at the other documents in a corpus, it seems a commonplace word. To skip the word, IDF (Inverse Document Frequency) which is estimated from DF (Document Frequency) is used. DF is the number of documents containing the word w_i which is appeared at documents in a corpus. The IDF is obtained by dividing the number of all documents by DF and then taking logarithm of that quotient as follows:

$$idf_i = \log \frac{|D|}{df_i}, \quad (2.2)$$

where $|D|$ is the total number of documents in the corpus and df_i is the number of documents where the word w_i appear. It is common to use $1 + df_i$, because if the word is not appeared in the corpus, this will lead to a division-by-zero. Finally, TF-IDF is denoted as follows:

$$tf-idf_{ij} = tf_{ij} \cdot idf_i . \quad (2.3)$$

TF-IDF weight gets higher when TF gets higher at a document and DF gets lower at all documents. Therefore, all words appears at the same time through the whole documents will be strained. For this reason, TF-IDF has frequently been used with Cosine Similarity to determine the similarity between two documents in vector space model.

2.3 Feature Extraction and Selection

Feature extraction is defined as transforming the input data into a reduced set of features when the input data to an algorithm is too huge to be processed and also there are extremely overlapping data i.e. much data, but not much information (Yang and Pedersen 1997). In text mining, usually noun as feature in documents is transformed into a vector. Recent works focused on using unigram and bigram to extract sentiments in English documents. Some researchers proved that unigrams show the best results in their experiment (Pang *et al.* 2002). In this case, a unigram is a part of n-gram which size is 1. An n-gram is a sub-sequence of n items from a given sequence (Christopher 1999). The items can be phonemes, syllables, letters, words or base pairs according to the application.

In Pang's case, unigram means a single word.

Feature selection has been grown in pattern recognition, statistics, and data mining field (El Alami 2009). The main idea is to select a subset of input variables by eliminating features with little or no predictive information. Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points (Kim 2003). Further, it is often the case that finding the correct subset of predictive features is an important issue in its own right. Feature selection in supervised learning has been well studied (Fukumizu 2003; Song *et al.* 2007), where the main goal is to find a feature subset that produces higher classification accuracy. Meanwhile feature extraction is to extract the meaningful set of features from the input data. It is used when the input data is too huge but there is no sufficient information that really has to be used. Each extracted words used to make a weight vector and the higher can be represented the better characteristics of the document. In information retrieval area, content words that extracted from the document are usually consisted of the noun and the verb. However, in sentiment classification problem, the adjective and the adverb are very important to extract features.

Now, I will briefly introduce some feature selection methods that used in this thesis to select proper features in general case.

2.3.1 Document Frequency

Document frequency is used in various fields as feature selection method. With ranked terms based on the frequency for each term t , this method is selecting the terms that are

most common in the class (Yang and Pedersen 1997). This is defined as follows:

$$DF(t) = df_t, \quad (2.4)$$

where df_t is same as mentioned in Section 2.2.

2.3.2 Chi-Square statistic

The Chi-Square (χ^2) statistic is the dependence between term t and class c (Galavotti *et al.* 2000). This is defined as follows:

$$\chi^2(t, c) = \frac{N \times (P(t, c) \times P(\bar{t}, \bar{c}) - P(\bar{t}, c) \times P(t, \bar{c}))^2}{P(t) \times P(\bar{t}) \times P(c) \times P(\bar{c})} \quad (2.5)$$

$$\chi^2(t) = \text{avg}_{i=1}^m \{ \chi^2(t, c_i) \}. \quad (2.6)$$

2.3.3 Information Gain

Information gain (Yang and Pedersen 1997) of a term measures the number of bits of information that obtained for category prediction by the presence or absence of the term in

a document. Let m be the number of class c . The information gain of a term t is defined as follows:

$$\begin{aligned} \text{IG}(t) = & -\sum_{i=1}^m P(c_i) \log P(c_i) + P(t) \sum_{i=1}^m P(c_i | t) \log P(c_i | t) \\ & + P(\bar{t}) \sum_{i=1}^m P(c_i | \bar{t}) \log P(c_i | \bar{t}). \end{aligned} \quad (2.7)$$

2.4 Classifiers

In this section, three classifiers used in this thesis will be shortly introduced: Naïve Bayes, k -NN, and SVM.

2.4.1 Naïve Bayesian Classifier

The Naïve Bayesian classifier (Mitchell 1997) is a classification algorithm based on Bayes rule that assumes the attributes x_1, \dots, x_n are all conditionally independent of one another, given y . The value of this assumption is that it dramatically simplifies the representation of $P(\mathbf{x}|y)$, and the problem of estimating it from the training data. Consider, for example, the case where $\mathbf{x} = (x_1, x_2)$. In this case,

$$\begin{aligned} P(\mathbf{x} | y) &= P(x_1, x_2 | y) \\ &= P(x_1 | x_2, y)P(x_2 | y) \\ &= P(x_1 | y)P(x_2 | y), \end{aligned} \quad (2.8)$$

where the second line follows from a general property of probabilities, and the third line follows directly from the above definition of conditional independence. More generally, when x contains n attributes which are conditionally independent of one another given y , it is defined as,

$$P(x_1, \dots, x_n | y) = \prod_{i=1}^n P(x_i | y). \quad (2.9)$$

2.4.2 k -Nearest Neighbor Algorithm

The k -Nearest Neighbor (k -NN) algorithm is a supervised learning algorithm where the result of new instance query is classified based on majority of k -nearest neighbor category (Mitchell 1997). The purpose of this algorithm is to classify objects based on closest attributes and training samples in the feature space. This algorithm is identified using a metric defined as below. Let x be an arbitrary instance with feature vector $\langle f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_n(\mathbf{x}) \rangle$ and Euclidean distance between two instances x_i and x_j is frequently used for real-valued features:

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^m (a_r(x_i) - a_r(x_j))^2}. \quad (2.10)$$

2.4.3 Support Vector Machine

Support vector machine (SVM) (Cortes and Vapnik 1995) is one of the supervised machine learning algorithms that can be used in classification or regression. SVM has been shown better results by several researchers on common text classification area (Joachims 1998; Tao *et al.* 2008) and it is also widely used in bioinformatics applications. Basically, the main idea is finding a hyperplane that has the largest distance between several classes in n -dimensional vector (a list of n numbers). In general, because SVM is a linear learning system that classifies two-class, SVM is effective machine learning algorithm to classify documents whether positive or negative. As an example, let the set of training data D be,

$$\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}, \quad (2.10)$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{in})$ is a n -dimensional input vector in a real-valued space $\mathbf{x}_i \subseteq \mathfrak{R}^n$. To build a classifier, SVM finds a linear function of the form is as follows,

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b, \quad (2.11)$$

where $f(\mathbf{x})$ is a real-valued function $f: \mathbf{X} \subseteq \mathfrak{R}$.

So, input vector \mathbf{x}_i is assigned to the positive class if $f(\mathbf{x}_i) \geq 0$, and to the negative class if $f(\mathbf{x}_i) < 0$.

$$y_i = \begin{cases} 1 & \text{if } (\mathbf{w} \cdot \mathbf{x}) + b \geq 0 \\ -1 & \text{if } (\mathbf{w} \cdot \mathbf{x}) + b < 0 \end{cases} \quad (2.12)$$

The class label y_i is either 1 or -1 that indicating the point \mathbf{x}_i belongs. 1 denotes the positive class and -1 denotes the negative class. The normal vector $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n) \in \mathfrak{R}^n$ is called weight vector. It is perpendicular to the hyperplane. The parameter $\frac{b}{\|\mathbf{w}\|}$ determines the offset of the hyperplane from the origin along the vector \mathbf{w} . $b \in \mathfrak{R}$ is called the bias. $\langle \mathbf{w} \cdot \mathbf{x} \rangle$ is the dot product of \mathbf{w} and \mathbf{x} .

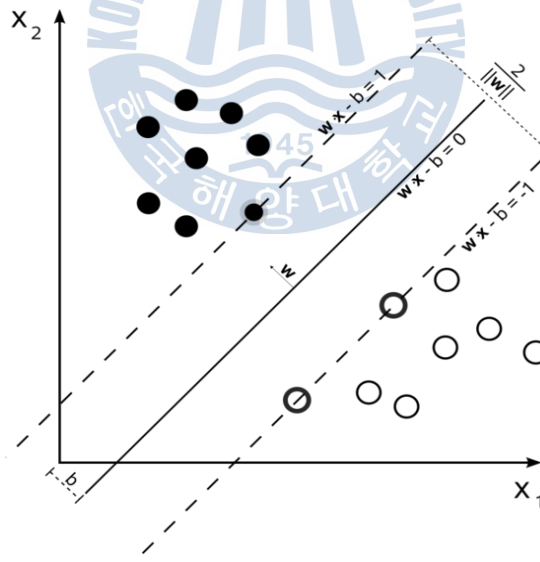


Figure 2.1: Maximum-margin hyperplane

Figure 2.1 describes an example of maximum margin for a SVM trained examples

from two classes and these margins are called the support vectors. In essence, SVM issues finding the hyperplane that satisfies $(\mathbf{w} \cdot \mathbf{x}) + \mathbf{b} = \mathbf{0}$, and this hyperplane is called the decision boundary or decision surface.

2.5 Accuracy Measure

In information retrieval fields, the precision is defined as the number of correct results that divided by the number of all returned results and the recall is defined as the number of correct results that divided by the number of results that should have been returned.

Table 2.1: Illustration of the classification context table

		Actual condition	
		Correct	Incorrect
Test (obtained) result	Positive	TP (true positive)	FP (false positive)
	Negative	FN (false negative)	TN (true negative)

Table 2.1 describes the terms of true positives, true negatives, false positives and false negatives. The precision and the recall are then denoted as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2.13)$$

$$\text{Recall} = \frac{TP}{TP + FN} . \quad (2.14)$$

The F_1 score (F-measure) is a popular measure that combines both the precision and the recall. The F_1 score can be interpreted as a weighted average of the precision and recall. In this thesis, the balanced F_1 score is denoted as follows,

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} . \quad (2.15)$$



Chapter 3

Feature Reweighting

This chapter mainly considers a novel feature reweighting method proposed by this thesis. At first usual way to extract features in Korean documents will be described and then based on this method, the feature reweighting method will be represented.

3.1 Feature extraction in Korean

To transfer a word in a document into a vector, word unigrams or bigrams as mentioned in Section 2.3 are used in general case of English. In Korean, however, they cannot be used directly because a basic Korean word has a lot of variants. Note that Korean is an agglutinative language. Furthermore, there are lots of difficult words like cants, argots, slangs or acronyms in most comments that particularly belong with news articles about major fields or politics. Most linguistic analyzers cannot understand them correctly and generate error messages or became to stop abnormally. Instead of word unigrams, the character bigrams or trigrams are used as features (i.e. 2 or 3 characters). Here are several reasons why in Korean this thesis uses character bigrams or trigrams:

- The length of about 80% words in character or syllable is 2 and 3 (Kim and Kim 2007).
- Character bigrams or trigrams can become sufficiently good features in the field of information retrieval (Lee *et al.* 1995; Jung 2004).
- Most comments in the news articles have lots of informal words such as cants, argots, slangs, acronyms and even emoticons, which can make errors in morphological analysis.

In this thesis, three types of text (i.e. body texts, its comments, and a sentiment lexicon) are transformed into bigrams or trigrams. After the transformation, features for machine learning algorithm are selected using feature selection methods mentioned in Section 2.3 (DF(t), χ^2 (t), IG(t)). Nevertheless useless words with high frequency are used as stopwords. The evaluation for those feature selection methods will be represented in Chapter 5.

3.2 Feature Reweighting Methods

The basic weighting scheme is TF-IDF as mentioned in Section 2.2. I represent a novel method for feature reweighting in sentiment classification of comments in a document, especially a Korean news article. The method adjusts the term frequency according to special conditions as in Equation (3.1) and (3.2).

$$tf'_{ij} = tf_{ij} + \alpha, \quad \alpha = \begin{cases} 2 & \text{if } t_{ij} \in B_{jk} \cap S \cap A_k \\ 1 & \text{if } t_{ij} \in B_{jk} \cap S \\ 0 & \text{otherwise} \end{cases} \quad (3.1)$$

$$tf''_{ij} = tf'_{ij} + \beta, \quad \beta = \begin{cases} \frac{|B_{jk} \cap S \cap A_k|}{2} & \text{if } t_{ij} = C_k \\ 0 & \text{otherwise} \end{cases}, \quad (3.2)$$

where the A_k denotes a feature set of words in the k -th body text, the B_{jk} denotes a feature set of words in the j -th comment on the k -th body text, the C_k denotes category information of the k -th body text, and the S denotes the sentiment lexicon. Basically the term frequency used in Equation (3.1) and (3.2) is same as the value that mentioned in Section 2.3. In Equation (3.1), the parameter α reinforces the term frequency of a term t_i which is included in the comment j , the sentiment lexicon, and the body text k . If the term t_i in comment j is included also in the sentiment lexicon, α should be a 1; $\alpha = 2$ if it is included in both the sentiment lexicon and the body text k . The category information (i.e. special category word) is added to the term frequency tf'_{ij} as the result of Equation (3.2). This will be explained more in next section with a factual example.

3.3 Example of Feature Reweighting Method

Figure 3.1 is an example of reweighting the parameter α and β . This figure describes

an example of the vector that consists of comment and category. As mentioned earlier, the value of α is determined according to the overlap of words in comments, body texts and the sentiment lexicon. The value of β is determined according to the category of the body text that has the comment.

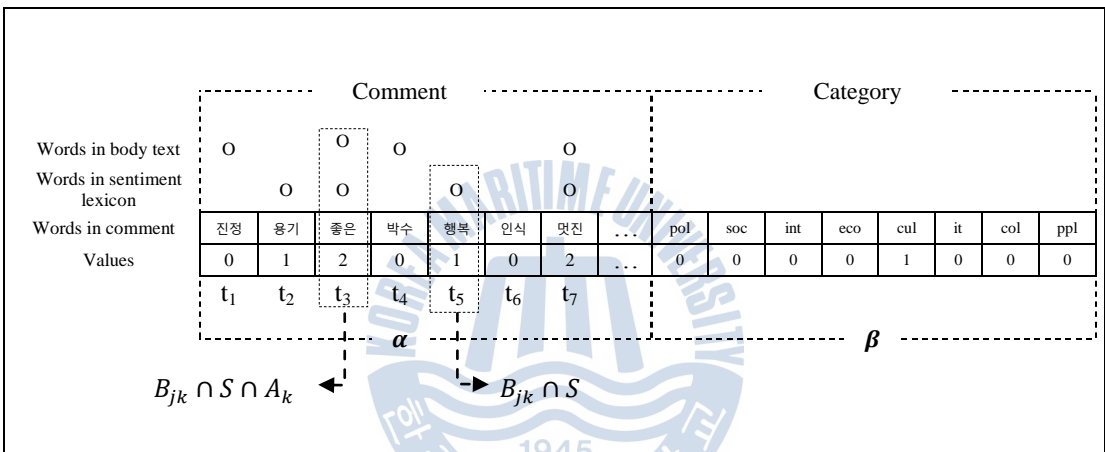


Figure 3.1: An example of feature reweighting for α and β

In Figure 3.1, the words t_3 and t_7 in the comment also exist in a sentiment lexicon and the body text, so its value of α is 2. In cases of the words t_2 and t_5 , α must be 1 because the words t_2 and t_5 exist also in the sentiment lexicon but not in the body text. These methods result from an importance of the body text and the sentiment lexicon. However, usually, there are lots of objective words (\neq sentiment word) in body texts. For this reason, not all words in the body text are helpful to reinforce weight of features in comments. Note that the words t_1 and t_4 get 0 (zero) because they exist in body texts but not in a sentiment

lexicon. After that, a value of β is determined according to the presence of category in the vector. If the category of the body text is determined, its value of $\frac{|B_{jk} \cap S \cap A_k|}{2}$ will be added.



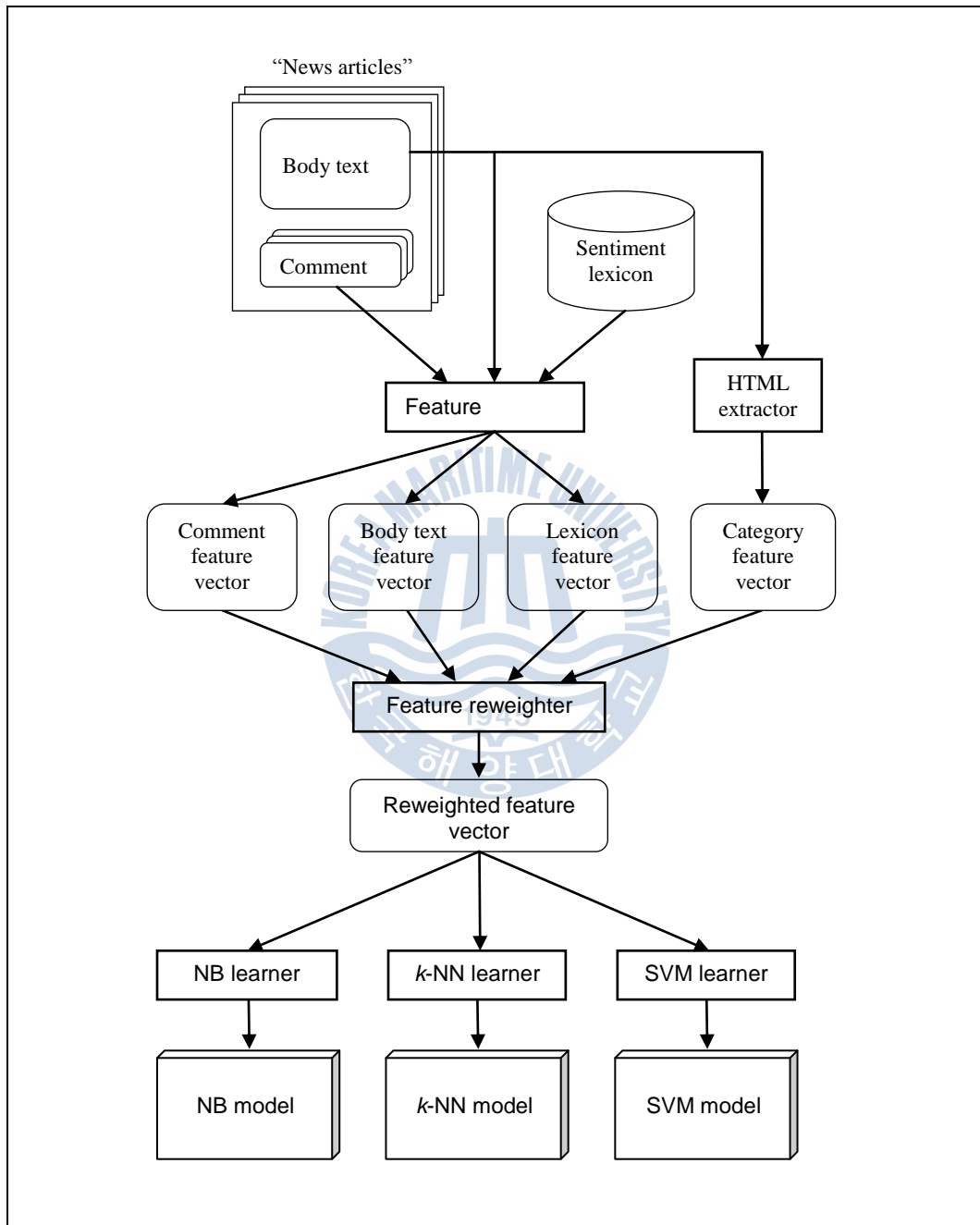
Chapter 4

Sentiment Polarity Classification System

This chapter describes the configuration for sentiment polarity classification that uses reweighted feature vector. The overall sentiment polarity classification method consists of a model generation part and a classification part. Both parts take a few documents such as body texts, comments and the sentiment lexicon unlike the other sentiment classification methods (Kennedy and Inkpen 2006; Pang *et al.* 2002). Three kinds of documents are to use them because each document has specific use for feature reweighting system.

4.1 Model Generation

First, the model generation system makes some models to use them where classification system should refer to. The models are made by three machine learning algorithms such as Naïve Bayes, k -NN, SVM. Figure 4.1 shows the architecture of model generation system.

**Figure 4.1: System architecture for the model generation**

This system consists of a feature extractor, an HTML extractor, a feature reweighter, and three learners. The feature extractor extracts proper features from a news article (a body text and its comments) and a sentiment lexicon by transforming them into character bigrams or trigrams. The reason why the documents are transformed into bigrams or trigrams was described in Section 3.1. The HTML extractor takes only a body text as an input, and determines which categories are the most proper to use it in feature reweighting phase. After each input data passes both extractors, feature vectors are made as a result of it. The feature reweighter calculates weight of vectors with these feature vectors using feature reweighting methods presented in Section 3.2. Then, reweighted feature vectors are presented from the feature reweighter. With these reweighted feature vectors, each learner generates the sentiment polarity classification models: NB model, k -NN model, SVM model. The sentiment polarity classification methods by using these models will be described at next section.

4.2 Sentiment Polarity Classification

The sentiment classification system takes a news article as an input, and then feature extractor extract proper features. This phase is same as model generation system. However classification system classifies category information from a body text instead of extracting it at HTML extractor. With various feature vectors from the feature extractor and category classifier, the feature reweighter generates a new reweighted feature vectors. Finally with reweighted vector, the sentiment polarity classifier presents a sentiment polarity as refers to the sentiment polarity classification models, i.e. NB model, k -NN model, SVM model. This system is described in Figure 4.2.

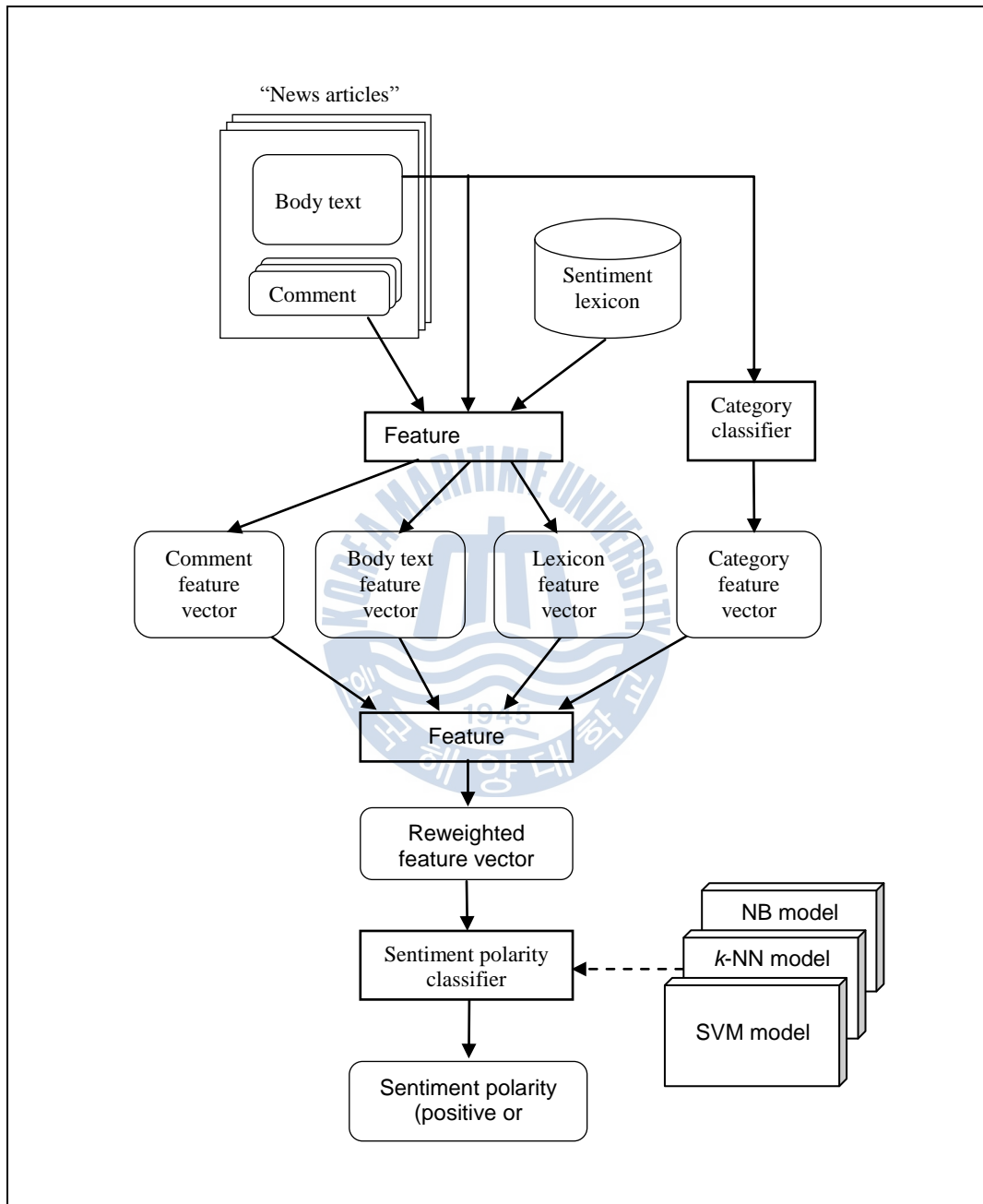


Figure 4.2: System architecture for the sentiment polarity classification

Chapter 5

Data Preparation

In this chapter, two kinds of data sets are described for Korean sentiment polarity classification. One is a training corpus for machine learning algorithms and the other one is a sentiment lexicon for feature reweighting. The two data sets are made directly because they are still not available publically in Korean. The method how to create them will be described in subsequent sections.

5.1 Korean Sentiment Corpus

To train and evaluate the proposed system for Korean sentiment classification, collected 1,377 news articles from one Korean news domain¹ and extracted html tags to classify category information in the body text are used. The category information is needed at reweighting phase and it is already described in Section 4.2. After extracting html tags, all body text and its comments are extracted from the each article. As it is mentioned in Section 2.3, both documents (comments and body texts) are divided into bi/trigram. Practically speaking, some documents without any comments are cast away because there are lots of useless contents such as advertisements or unrecognizable words in some comments. As a result, 8,320 comments are collected. Table 5.1 shows the statistics on collected documents as an evaluation corpus.

¹ <http://www.donga.com>

Table 5.1: The statistics on constructed corpus

	Body texts	Comments
Total number	1,377	8,320
Average number of comments	-	6
Number of words	863,379	274,626
Average number of words	627	33

In Table 5.1, the average number of comments per document is about 6 and a document has sufficient comments to evaluate the proposed system. The average number of words per comments is 33 and they consist of 2 or 3 sentences on average. Therefore a comment can be considered as a very short document. Each comment is annotated with the polarity judgments (i.e. positive and negative) manually. As a result, finally 7,511 negative comments and 809 positive comments are collected.

The body texts as documents have their own categories as mentioned before and one or more comments. This thesis uses 8 categories: Politics, Economy, International, Society, IT, Column, People, and Culture. Table 5.2 shows the number of documents (body texts and comments) which will be used in the experiment.

Table 5.2: Categories of documents

Category	Number of body texts	Number of comments	Number of positive comments (%)	Number of negative comments (%)
Politics	462	4414	306 (6.9%)	4108 (93.1%)
Society	321	2085	234 (11.2%)	1851 (88.8%)
International	107	299	37 (12.4%)	262 (87.6%)
Economy	293	985	118 (12.0%)	867 (88.0%)
Culture	124	322	54 (16.8%)	268 (83.2%)
IT	13	27	7 (25.9%)	20 (74.1%)
Column	44	117	32 (27.4%)	85 (72.6%)
People	13	33	17 (51.5%)	16 (48.5%)
Total	1377	8320	809 (9.7%)	7511 (90.3%)

In Table 5.2, negative comments accounts for a large proportion of news articles. Especially the proportion of negative comments in politics articles is the largest as assumed before.

5.2 Korean Sentiment Lexicon

Unfortunately there's no public Korean sentiment lexicon. For this reason, a Korean sentiment lexicon should be built directly. This thesis used a Korean dictionary to build a Korean sentiment lexicon. Just extracting Korean sentiment words from a Korean dictionary needs huge man power and lots of time. Therefore this thesis used an *English subjectivity lexicon*² which is publically made earlier to build the Korean sentiment lexicon.

² <http://www.cs.pitt.edu/mpqa/>

As a result of this task, all words that gathered from the English subjectivity lexicon are 4,138 negative English words and 2,297 positive English words. Then, English words in the lexicon are translated into corresponding Korea words by using an English-Korean dictionary semi-automatically as a primary Korean sentiment lexicon. The Korean sentiment lexicon is expanded by appending synonyms and antonyms through a Korean dictionary. After expanding the Korean sentiment lexicon, all overlapped words and meaningless words in Korean are eliminated. Finally, the Korean sentiment lexicon consists of 4,046 negative words and 3,044 positive words. The lexicon is involving nouns, adjectives and adverbs. Table 5.3 shows the number of total sentiment words used for the proposed system.

Table 5.3: The number of total sentiment words

Polarity	English	Korean
Negative	4,138	4,046
Positive	2,297	3,044

Chapter 6

Experiments

Now there are three questions to be answered in the experiments:

- (1) Which kind of a classifier is most appropriate for sentiment classification for comments in Korean?
- (2) Are body texts helpful for identifying the sentiment polarity of its comments?
- (3) Are character n -grams sufficient for sentiment polarity classification for comments in Korean?

Before dealing with the problems, the environment of the experiments has to be discussed. Next section treats the experimental environment.

6.1 Experimental Environment

The corpus described in Chapter 5 is used to evaluate the method mentioned earlier. The corpus involves 8,320 comments, and this is not sufficient for sentiment polarity classification. Therefore this thesis uses a cross-validation technique (Kohavi 2005): every evaluation has been performed by 4-fold cross validation (for example, using training data: 6,240, test data: 2,080). Cross-validation usually involves splitting a data set into two

where one piece of it is used to train up a model (for example using SVM) and the other piece of the data is used to test or evaluate that model. The macro-averaging³ means that calculating for each category first and then averaging them. In the other hand, the micro-averaging means that calculating over all decisions and then averaging them. The two procedures bias the results differently micro averaging tends to over-emphasize the performance on the largest categories, while macro-averaging over-emphasizes the performance on the smallest. In the thesis, only macro-averaging method is used because there are just two categories (positive vs. negative).

In the public domain, there are a lot of machine learning tools such as Weka⁴ and AI::Categorizer⁵. The AI::Categorizer which is a framework for automatic text categorization is used in this thesis and it consists of a collection of Perl modules that implement common categorization tasks.

6.2 Experimental Results

In this section, the four types of features that described in Chapter 3 and 4 with SVM will be evaluated:

Type 1: A (words in a comment)

Type 2: A + S (words in the sentiment lexicon)

Type 3: A + S + B (words in the body text)

Type 4: A + S + B + C (a category of the body text)

³ <http://backpan.perl.org/authors/id/K/KW/KWILLIAMS/Statistics-Contingency-0.02.readme>

⁴ <http://www.cs.waikato.ac.nz/ml/weka/>

⁵ <http://search.cpan.org/~kwilliams/AI-Categorizer-0.09/>

This thesis presents a pertinent feature reweighting scheme according to feature types as it is mentioned in Section 3.2. In case of the feature type 1, any feature reweighting scheme is not used and it is defined as a basic model. In case of the feature types 2 and 3, the feature reweighting schemes Equation (3.1) is applied: each parameter α in Equation (3) is 1 and 2. In case of the feature type 4, all features in the feature type 3 are included, also with the category information described in Equation (3.2). All evaluation of these feature reweighting methods will be presented below.

6.2.1 Classifiers

This section is going to answer the first question, that is, “which kind of classifier is most appropriate for sentiment classification for comments in Korean?” In general, there are lots of parameters which influence the performance of classifiers.

One of Naïve Bayes, k -NN, and SVM classifier is a candidate of the most proper classifier. Figure 6.1 shows the macro-averaged performance of each classifier and all features consist of bigram.

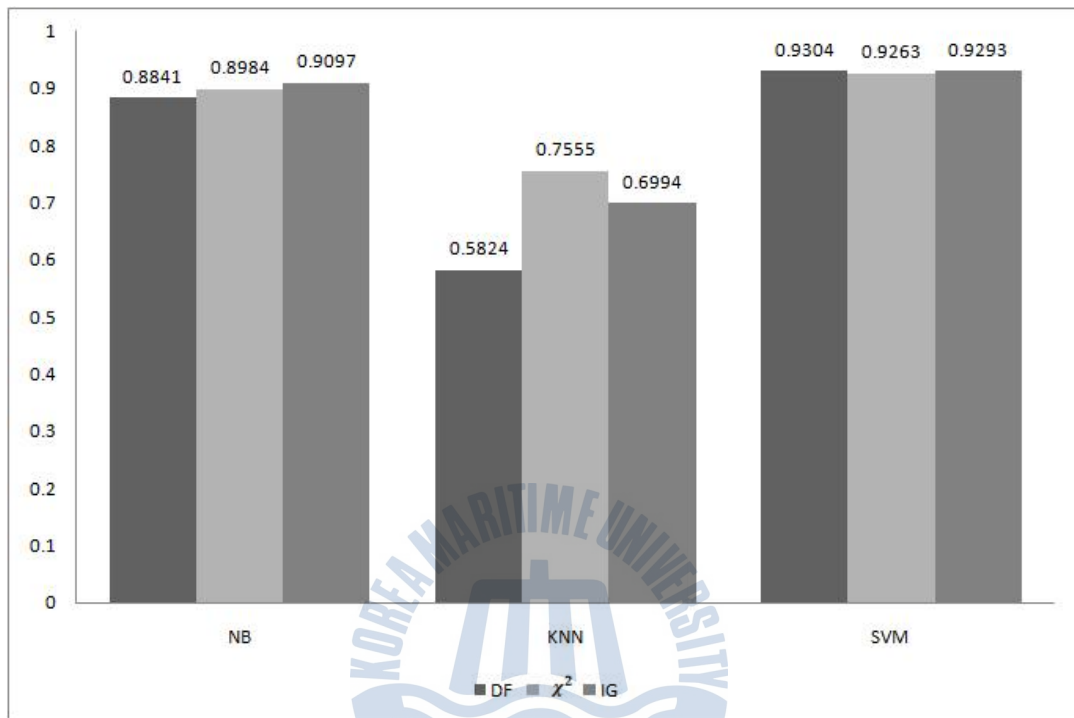


Figure 6.1: Performance according to classifiers

In general, the SVM had shown better performance than the other learner in text classification (Manning *et al.* 2008). Figure 6.1 shows the same result with it. This result is about the average F_1 score for 50-1,000 selected features. For this reason, SVM is decided as a default classifier.

6.2.2 Features on Body Text

At first, Figure 5.2 shows the overall F_1 score when 50 – 1,000 features are selected by

calculating document frequency. According to the result, answers for the second and the last question which are mentioned in Chapter 5 can be found. That is, body text is helpful to identify the polarity of its comments. Also the result shows that n-gram is sufficient to classify sentiment of document in Korean. Next figure describes average F_1 score when use SVM as a default classifier.

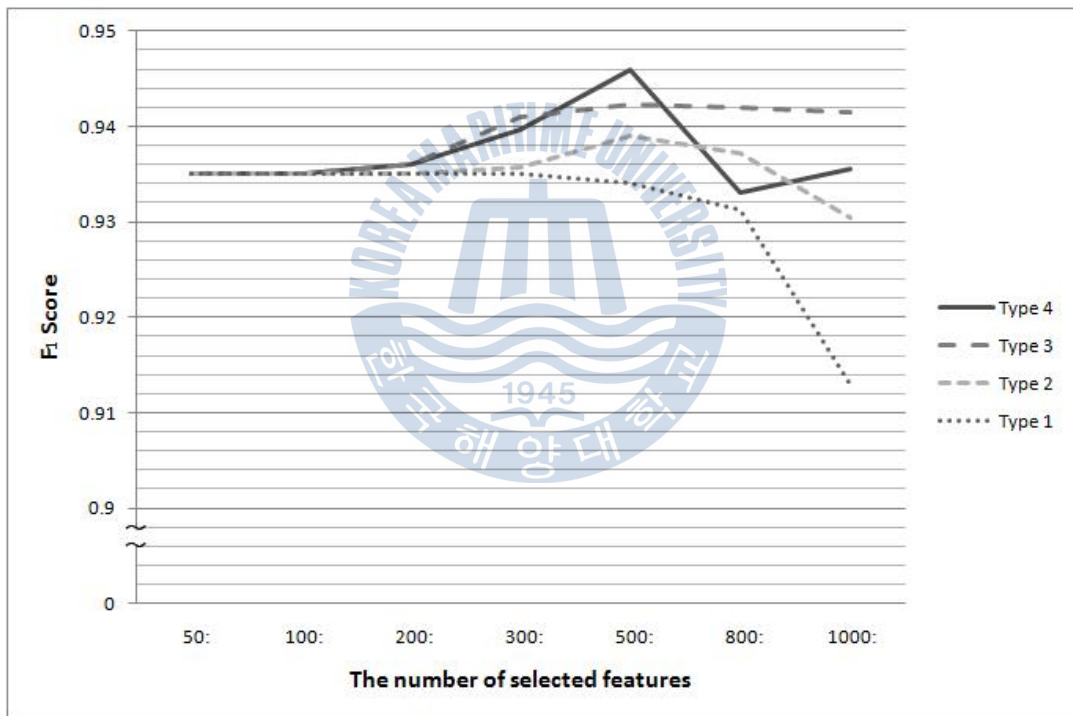


Figure 6.2: Overall F_1 score (Document Frequency, SVM)

Using document frequency to select proper features for the all feature reweighting methods that introduced at the top part of this chapter does not always shows the best

performance. However, in case of using χ^2 statistic and Information Gain, there are different aspects from Document Frequency. Figure 6.2 and 6.3 describe the overall F_1 score at the same environment with Figure 6.2, but they use χ^2 statistic or Information Gain to select proper features for the sentiment classifier.

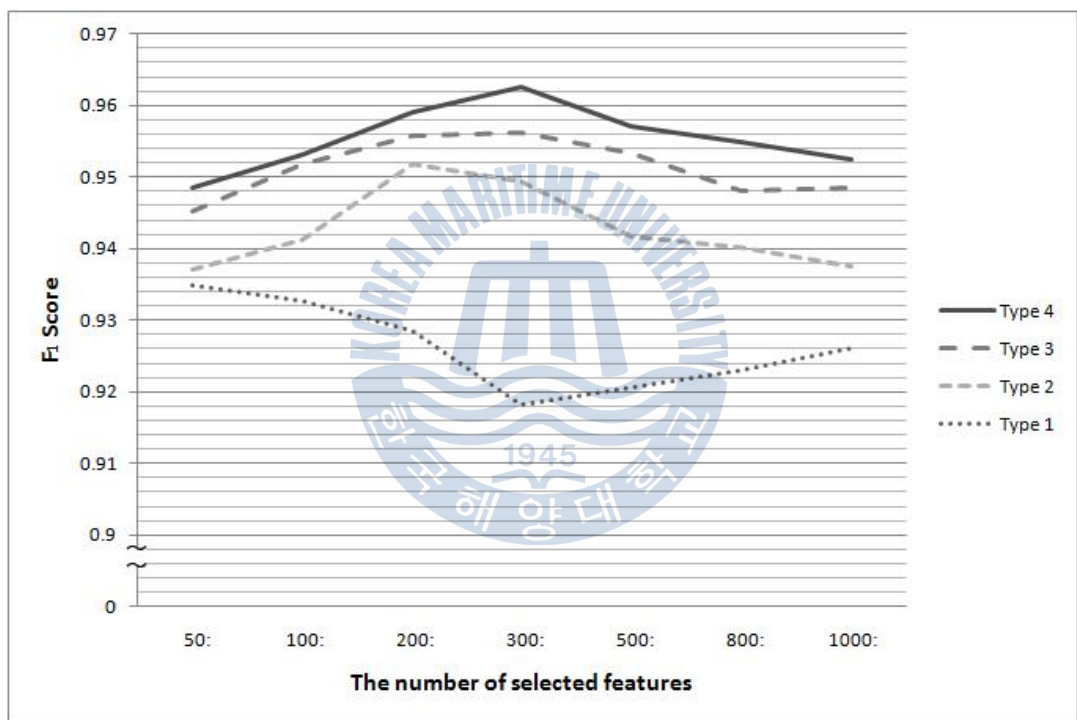


Figure 6.3: Overall F_1 score (χ^2 statistic, SVM)

In Figure 6.3, using χ^2 statistic has shown comparatively equal growth between reweighting steps. And using Information has shown similar results either as follows.



Figure 6.4: Overall F_1 score (Information Gain, SVM)

The case (4) dropped under the case (3) at the point of features 100 and 800, but generally using all features (i.e. the case 4) have shown better performance than using sentiment lexicon and words in the article (i.e. the case 3). Also in both Figure 6.3 and 6.4, using feature reweighting methods (i.e. adding features in a sentiment lexicon and a body text) has grown up.

6.2.3 Character n-Grams in Korean

To see more about the growth, the result for both bigram and trigram using χ^2 statistic with 300 selected features is shown at Figure 6.5.

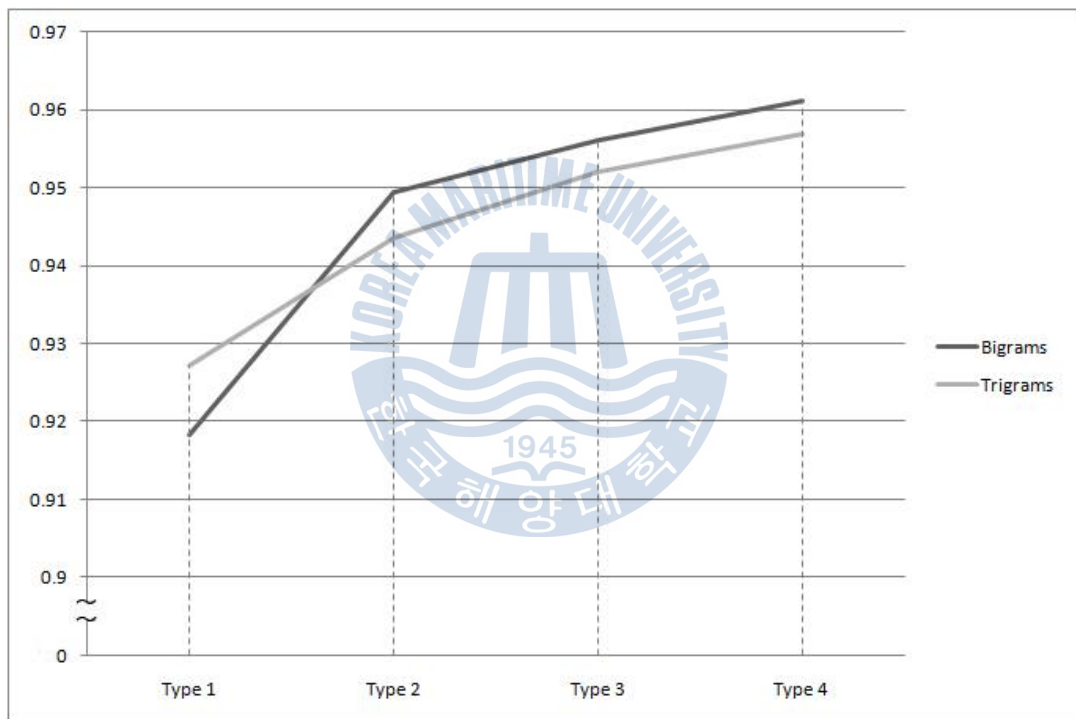


Figure 6.5: The result of F_1 score (χ^2 statistic, SVM, 300 features)

With these facts, feature-adjusting methods can be the answer for the question (2) and (3) that mentioned in Chapter 6. These results are caused by the characteristic of feature selection methods. The χ^2 statistic and the information gain are relative to a class or classes

of document but document frequency is a one of class-independent measures (Yang and Pedersen 1997). That is, using document frequency just deals with terms in a comment to select features without the association between class and word. However χ^2 statistic and information gain try to select proper words with association between a class and a word.



Chapter 7

Conclusions and Future Works

This thesis presents feature reweighting methods for sentiment polarity classification of comments in Korean news articles using machine learning. Proposed feature reweighting method needs a Korean sentiment corpus but it is not available yet. Thus, the corpus that consists of 1,377 body texts and 8,320 comments from Korean news articles is used. The method uses specific feature sets which are a sentiment lexicon, feature sets related with body texts in news articles, and category information for the article. The Korean sentiment lexicon is not available either, so it is built from an English sentiment lexicon as using an English-Korean dictionary. To evaluate the method, several classifiers i.e. Naïve Bayes, k -NN, SVM with three feature selection methods i.e. Document Frequency, χ^2 statistic, Information Gain are examined and conclude related results. Finally, this thesis has demonstrated that sentiment words and body text are effective for sentiment polarity identification of comments in news articles. Each document enhances the F_1 score. However not all element could help to enhance the performance. As an example, some words in a body text are meaningless in sentiment. It means that a word without sentimental sense is hard to give great effect to raise its value. Therefore choosing words that have a sentimental meaning is important to select proper features.

In future work, I will test with other feature selection methods for the best performance on each classifier. Also I will deal with syntactic or semantic processing such as negation problems. As an example, the sentence “You’ll never be disappointed.” has to be

processed before selecting features. Also I will adapt the method that distinguishes the comments for other subjects such as review sites or private blogs.



Bibliography

- Cao, Y., Xu, J., Liu, T. Y., Li, H., Huang, Y., and Hon, H. W. 2006. "Adapting ranking SVM to document retrieval". *Proceedings of the annual Association for Computing Machinery Special Interest Group on Information Retrieval Conference on Research and Development in Information Retrieval*, pp. 186-193.
- Cortes, C. and Vapnik, V. 1995. "Support-vector networks". *Machine Learning*, pp. 273-297.
- Manning, C. D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.
- Devitt, A. and Ahmad, A. 2007. "Sentiment polarity identification in financial news: A cohesion-based approach". *Proceedings of the Association for Computational Linguistics*, pp. 984-991.
- El Alamia, M. E. 2009. "A filter model for feature subset selection based on genetic algorithm". *Knowledge-Based Systems*, vol. 22, no. 5, pp. 356-362.
- Falquet, G., Guyot, J., Nerima, L., and Simi, M. 2000. "Feature selection and negative evidence in automated text categorization". *Proceedings of the European Conference on Research and Advanced Technology for Digital Libraries*, pp. 59-68.
- Fukumizu, K., Bach, F., and Jordan, M. 2003. "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces". *The Journal of Machine Learning Research*, vol. 5, pp. 73-99.
- Hatzivassiloglou, V. and McKeown, K. 1997. "Predicting the semantic orientation of adjectives". *Proceedings of the European Chapter Meeting of the Association for Computational Linguistics*, pp. 174-181.

- Joachims, T. 1998. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features". *Proceedings of the European Conference on Machine Learning*, pp. 137-142.
- Jung, C. 2004. *An Indexing Method Based on the Mixed n-gram for Korean Information Retrieval*. Master Thesis, Department of Computer Engineering, Korea Maritime University.
- Karen, J. S. 1972. "A statistical interpretation of term specificity and its application in retrieval". *The Journal of Documentation*, vol. 28, no. 1, pp. 11-21.
- Kennedy, A. and Inkpen, D. 2006. "Sentiment classification of movie reviews using contextual valence shifters". *Computational Intelligence*, vol. 22, no. 2, pp. 110-125.
- Kim, C. and Kim, Y. 2007. "Statistical information of Korean dictionary to construct an enormous electronic dictionary". *The Journal of Korean Contents Society*, vol. 7, no. 6, pp. 60-68.
- Kim, J., et al. 1993. "Korean part-of-speech tagging by using a fuzzy net". *The Conference on Hangul and Korean Language Information Processing*, pp. 593-603.
- Kim, S. and Hovy, E. 2004. "Determining the sentiment of opinions". *Proceedings of the International Conference on Computational Linguistics*, pp. 1367-1373.
- Kim, S. and Hovy, E. 2005. "Automatic detection of opinion bearing words and sentences". *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 61-66.
- Kim, Y., Street, W. N., Nick, W., and Menczer, F. 2003. Feature Selection in Data Mining. *Data Mining: Opportunities and Challenges*, pp. 80-105.
- Kohavi, R. 2004. "A study of cross-validation and bootstrap for accuracy estimation and model selection". *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1137-1143.

- Lee, J., Park, H., Park, H., Ahn J., and Kim, M. 1995. "An effective indexing methods for Korean text". *Proceedings of the Korean Society for Information Management Conference*, pp. 11-14.
- Li, T., Zhu, S., and Ogihara, M. 2008. "Text categorization via generalized discriminant analysis". *Information Processing & Management: an International Journal*, vol. 44, no. 5, pp. 1684-1697.
- Manning, C., Raghavan, P., and Schütze H. 2009. *Introduction to Information Retrieval*. Cambridge University Press.
- Miller, G.A. 1990. "Nouns in WordNet: A lexical inheritance system". *International Journal of Lexicography*, vol. 3, no. 4, pp. 245-264.
- Mitchell, T. 1997. *Machine Learning*. McGraw Hill.
- Pang, B. and Lee, L. 2004. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts". *Proceedings of the Association for Computational Linguistics*, pp. 271-278.
- Pang, B. and Lee, L. 2008. "Opinion mining and sentiment analysis". *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1-135.
- Pang, B., Lee, L., and Vaithyanathan, S. 2002. "Thumbs up? Sentiment classification using machine learning techniques". *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 79-86.
- Riloff, E. and Wiebe, J. 2003. "Learning extraction patterns for subjective expressions". *Proceedings of the Conference on Empirical methods in Natural Language Processing*, vol. 10, pp. 105-122.
- Riloff, E., Wiebe, J., and Wilson, T. 2003. "Learning subjective nouns using extraction pattern bootstrapping". *Proceedings of the Conference on Computational Natural Language Learning*, vol. 4, pp. 25-32.

- Rimon, M. 2005. "Sentiment classification: Linguistic and non-linguistic issues". *Proceedings of Israel Association for Theoretical Linguistics 21*, <http://linguistics.huji.ac.il/IATL/21/>.
- Salton, G., Wong, A., and Yang, C. S. 1975. "A vector space model for automatic indexing". *Communications of the Association for Computing Machinery*, vol. 18, no. 11, pp. 613-620.
- Song, L., Smola, A., Gretton, A., Borgwardt, K., and Bedo, J. 2007. "Supervised feature selection via dependence estimation". *Proceedings of the International Conference on Machine Learning*, pp. 82 -830.
- Spertus, E. 1997. "Smokey: Automatic recognition of hostile messages". *Proceedings of the International Conference on Innovative Application of Artificial Intelligence*, pp. 1058- 1065.
- Turney, P. 2002. "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". *Proceedings of Association for Computational Linguistics*, pp. 417-424.
- Whissell, C. M. 1989. "The dictionary of affect in language". In Plutchik R. and Kellerman H. (Ed.), *Emotion: Theory, Research, and Experience*. New York: Academic Press, pp. 113-131.
- Wiebe, J., Bruce, R., and O'Hara, T. 1999. "Development and use of a gold standard data set for subjectivity classifications". *Proceedings of the Association for Computational Linguistics*, pp. 246-253.
- Wiebe, J. 2000. "Learning subjective adjectives from corpora". *Proceedings of the Association for the Advanced of Artificial Intelligence*, pp. 735-740.
- Wiebe, J., Bruce, R., and Bell, M. 2001. "Identifying collocations for recognizing opinions". *Proceedings of the Association for Computational Linguistics Workshop on Collocation: Computational Extraction, Analysis and Exploitation*, pp. 24-31.

- Wilson, T., Wiebe, J., and Hoffmann, P. 2005. "Recognizing contextual polarity in phrase-level sentiment analysis". *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pp. 347-354.
- Yang, Y. and Pedersen, J. O. 2007. "A comparative study on feature selection in text categorization". *Proceedings of the International Conference on Machine Learning*, pp. 412-420.



Acknowledgements

First of all, I would like to thank God for everything because I'm nothing without him.

I specially want to thank my supervisor Jae-Hoon Kim for the great support he has been continuously offering to me, and for being there whenever I needed him, patiently listening to my anxieties and queries.

Thanks also to Prof. Gil-Su Ryu and Prof. Hyu-Chan Park for their thorough reviewing of this thesis and their valuable comments.

The Department of Computer Engineering and Natural Language Processing Laboratory have provided an excellent environment for my research. I spent many enjoyable hours with department and laboratory members. Without this rich environment I doubt that many of my ideas would have come to fruition.

There are also a number of persons, who may not have been directly involved in this thesis, but without whom things would have been much harder. I thank from the bottom of my heart to my lovely people, Eric Connelly, Justin Steffen, Yuki Ikeda, including "Korean family" too many individually to mention, for being there whenever I needed you. I will never forget the tough and good times we shared in Champaign-Urbana, USA.

I would like to thank also to my true friends, Dae-Sung Park, Hyun-Hee Lee, Seong-Hwan Jeon, Tae-Jin Kim and Kyung-Kook Kim for the great support to me, and trust me whatever I did.

Finally, most special thanks go to my family who has provided me with continual support and love, guidance. They constantly remind me of their confidence and encourage me whenever I am in doubt.

I hope that all my lovely people are always happy, and will have a peaceful life in Jesus.



A THESIS
SUBMITTED IN PARTIAL FULFILLMENT OF THE THESIS
REQUIREMENTS FOR THE DEGREE OF MASTER OF COMPUTER
ENGINEERING

KOREA MARITIME UNIVERSITY

© Copyright by Hyung-Won Seo 2009
All Rights Reserved