JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice

Research Paper
J. of infosci. theory and practice 1(1): 10-26, 2013
http://dx.doi.org/10.1633/JISTaP.2013.1.1.1

Domain Adaptation for Opinion Classification: A Self-Training Approach

Ning Yu*

School of Library and Information Science
University of Kentucky, USA
E-mail: ning.yu@uky.edu

ABSTRACT

Domain transfer is a widely recognized problem for machine learning algorithms because models built upon one data domain generally do not perform well in another data domain. This is especially a challenge for tasks such as opinion classification, which often has to deal with insufficient quantities of labeled data. This study investigates the feasibility of self-training in dealing with the domain transfer problem in opinion classification via leveraging labeled data in non-target data domain(s) and unlabeled data in the target-domain. Specifically, self-training is evaluated for effectiveness in sparse data situations and feasibility for domain adaptation in opinion classification. Three types of Web content are tested: edited news articles, semi-structured movie reviews, and the informal and unstructured content of the blogosphere. Findings of this study suggest that, when there are limited labeled data, self-training is a promising approach for opinion classification, although the contributions vary across data domains. Significant improvement was demonstrated for the most challenging data domain—the blogosphere—when a domain transfer-based self-training strategy was implemented.

Keywords: Domain adaptation, Opinion classification, Self-training, Semi-supervised learning, Sentiment analysis, Machine learning

1. INTRODUCTION

The rapid growth of freely accessible and easily customizable Web 2.0 applications has made it easy and fun for people to share their experiences, knowledge, and opinions. Retail websites such as Amazon.com and review aggregators such as Yelp.com collect customer reviews on specific products or services while blogs and social networking sites such as Twitter and Facebook allow users to publish opini-

Open Access

Received date: December 30, 2012
Accepted date: February 23, 2013

***Corresponding Author:** Ning Yu
Assistant professor
School of Library and Information Science
University of Kentucky, USA
E-mail: ning.yu@uky.edu

ons and share emotions on an infinite array of topics ranging from the benefits of eating blueberries to the U.S. presidential election. Being able to listen to and understand online voices is playing an important role in today's decision making for business practices, political campaigns, and daily life.

Since the late 1990s, researchers from different communities have been working in the area of sentiment analysis, which includes tasks such as differentiating opinions from facts (Wiebe, Wilson, Bruce, Bell, & Martin, 2004; Yang, Yu, & Zhang, 2007), detecting positive and negative polarity (Abbasi, Chen, & Salem, 2008; Pang, Lee, & Vaithyanathan, 2002), classifying fine-grain emotions (Bollen, Mao & Zeng, 2011; Yu, Kübler, Herring, Hsu, Israel & Smiley, 2012), and identifying other opinion properties (Tsou, Yuen, Kwong, Lai, & Wong, 2005; Ku & Chen, 2007). For any tasks, data pre-labeled with sentiment categories are essential for creating and evaluating sentiment analysis systems. However, the reality is that labeled data are usually limited, especially at the sub-document level. Although this shortage of sentiment-labeled data is less challenging in some domains (e.g., movie reviews) than in others (e.g., blog posts), simply borrowing labeled data from a non-target data domain often fails due to the domain transfer problem.

Domain transfer is a widely recognized problem for machine learning algorithms because models built via learning one data domain generally do not perform well in another data domain. Hence for each data domain, machine learning tends to start from scratch. But there may not be sufficient 'ground truth' (i.e., labeled data) in the target data domain for machine learning algorithms to rely on. While it is difficult to obtain sentiment-labeled data and manual annotation is tedious, expensive, and error-prone, unlabeled user-generated data are readily available. This paper therefore examines strategies to utilize both unlabeled data in the target domain and labeled data in other data domains to tackle the domain transfer problem. The specific machine learning methods explored in this research fall into the category of semi-supervised learning (SSL), which requires only limited labeled data to automatically label unlabeled data. SSL has achieved promising results in sparse data situations in various natural language processing (NLP) tasks, including topic classification and sentiment analysis; but SSL has seldom been examined for domain adaption.

Specifically, this study investigates applications of self-training for opinion classification in three types of Web content: edited news articles, semi-structured movie reviews, and the informal and unstructured content of the blogosphere. An easily generalizable and highly adaptable SSL algorithm, self-training, is evaluated for its effectiveness in sparse data situations and domain adaptation.

2. BACKGROUND AND RELATED WORK

Two major sentiment analysis strategies exist in the sentiment analysis literature: The ad hoc rule-based approach, sometimes known as the lexicon-based approach (Ounis, Macdonald, & Soboroff, 2008), and the machine learning-based approach, sometimes known as the corpus-based approach. Both of these approaches benefit from the large number and great variety of sentiment-bearing features used as evidence in sentiment analysis. Such sentiment evidence can be knowledge-based (e.g., the more depressed a person feels, the more likely he/she will use the first-person word "I," Pennebaker, 2011), statistical/empirical (e.g., high order n-grams), or style-based (e.g., "IMHO," "-)"). Since each source of sentiment evidence has its own characteristics and captures different aspects of sentiment, sentiment-bearing features from more than one source of evidence are often preferred. Most studies have suggested that a fusion of various sentiment-bearing features surpasses the use of any single subset of features (Chesley, Vincent, Xu & Srihari, 2006; Gamon, 2004; Hatzivassiloglou & Wiebe, 2000; Yang, et. al., 2007).

The machine learning approach is more practical in sentiment analysis than the ad hoc rule-based approach due to its fully automatic implementation and its ability to identify features that are not intuitive to human. State-of-the-art topical supervised classification algorithms are often tailored for sentiment analysis in the following manner: 1) binary feature values (presence/absence) are used instead of frequency. This is motivated by the extreme

brevity of the classification unit (e.g., tweets, reviews) and the characteristics of sentiment analysis, where occurrence frequency is less influential (i.e., a single occurrence of sentiment evidence is sufficient); and 2) a wider variety of evidence (e.g., linguistic features, links) is investigated in addition to auto-generated features (e.g., bag-of-words, n-grams). These supervised learning algorithms have achieved satisfactory results for sentiment analysis (Wiebe, et. al., 2004; Zhang & Yu, 2007).

The biggest limitation associated with supervised learning is that it is sensitive to the quantity and quality of the training data and may fail when training data are biased or insufficient. In contrast with supervised learning, which learns from labeled data only, *semi-supervised learning* (SSL) learns from both labeled and unlabeled data based on the idea that although unlabeled data hold no information about classes (e.g., "sentiment" and "non-sentiment"), they do contain information about joint distribution over classification features. In contrast with supervised learning, the value of SSL in sentiment analysis lies not only in its need for less labeled data but also in its ability to handle the domain dependency challenge: When there are labeled data in the non-target data domain only, an SSL algorithm can reduce the bias of the non-target data by increasing the number of labeled data from the target data domain. This aspect of SSL is very attractive for sentiment analysis in challenging data domains such as the blogosphere, which is short of high-quality sentiment-labeled data.

2.1. Semi-Supervised Learning and Self-Training

According to a survey of SSL by Zhu (2008), the most commonly used SSL algorithms include self-training, Expectation-Maximization (EM) with generative mixture models, co-training, Semi-Supervised Support Vector Machines (S^3VMs), and graph-based methods. Except for S^3VMs, all SSL algorithms have been found to be effective for sentiment analysis (Aue & Gamon, 2005; Pang & Lee, 2004; Yu & Kübler 2010; 2011). This study focuses on self-training due to its easy generalization and high adaptability.

Self-training[1] is a wrapper SSL approach that can be applied to any existing system as long as a confidence score can be produced. Self-training keeps a system in a black box and avoids dealing with any inner complexities. The major steps in self-training are: (1) training an initial classifier on the labeled dataset; (2) applying this classifier to the unlabeled data and selecting the most confidently labeled data as determined by the classifier to augment the original labeled dataset; and (3) re-training the classifier by repeating the whole process from step (1). A simple pseudo code for self-training is illustrated in Figure 1.

Input: classifier *C*, a small set of labeled data *L*, a large amount of unlabeled data *U*

Loop until iteration procedure converges or loop for *k* iterations

1) Train *C* on *L;*
2) Apply trained *C* to *U;*
3) Select top *n* results returned by *C* and add them back to *L;*

Output: Extended *L* and an updated *C*

Fig. 1 Bootstrapping Procedure in Self-training

[1] Self-training, also known as mutual bootstrapping or self-teaching, is conceptually equal to the pseudo relevance feedback technique in information retrieval where the top n retrieved results to a given query are assumed to be relevant and are used to form a new query.

Self-training has been originally adopted for sentiment lexicon expansion (Riloff & Jones, 1999) and only recently has been explicitly applied for sentence-level sentiment analysis (He & Zhou, 2011). The initial classifier C is either a simple rule-based classifier built using a few manually created opinion seeds or a supervised classifier trained on a few manually labeled data. Across several experiments carried out by Wiebe and Riloff (2005), a self-trained Naïve Bayes classifier using this procedure achieved the best recall with modest precision when classifying subjective sentences.

2.2. Domain Adaptation

Domain dependency may seem less problematic for sentiment analysis than topical classification since generic sentiment-bearing words such as "good" and "bad" are not limited to any particular domain. But there are few generic sentiment-bearing words and it is therefore necessary to extract sentiment-bearing features from the target data collection. These features are generally domain dependent and may not be reusable in another domain for several reasons: (1) there are specific sentiment-bearing words associated with different domains (e.g., "cheap" and "long-lasting" are frequently used in product reviews, but not in movie reviews); (2) different domains have different stylistic expectations for language use (e.g., news articles are less likely than blogs to use words such as "crappy" or "sooooooooo"); and (3) some sentiment-bearing words can be either positive or negative depending on the object (e.g., "small" may be positive in "a small camera" but negative in "a small memory card"). Since information used for sentiment analysis is typically lexical and lexical means of expressing sentiments may vary not only from domain to domain but also from register to register, a sentiment analysis strategy that works for one target data domain generally will not work for another data domain.

Most sentiment analysis systems borrow sentiment-labeled data directly from non-target data domains when there are few labeled data in the target domain or when the characteristics of the target domain make it difficult to detect sentiments if the non-target data appear to be "relevant to the application and of sufficient quantity" (Conrad & Schider, 2007, p. 235). This approach is especially common in opinion detection in the blogosphere. For example, Chesley et al. (2006) leveraged blog training data with non-blog training data containing relatively "pure" opinion information; additionally, most participants in TREC's Blog track have crawled the Web to generate a great number of opinion-labeled training data. However, according to Aue and Gamon (2005), who compared four strategies for utilizing opinion-labeled data from one or more non-target domains, using non-target labeled data without an adaptation strategy is not as efficient as using labeled data from the target domain, even when the majority of labels are assigned automatically by a self-training algorithm.

Blitzer, Dredze and Pereira (2007) proposed a structural correspondence learning (SCL) algorithm for sentiment classification to reduce the classification error of a classifier trained with non-target data. The key to this domain adaptation strategy is to implicitly associate domain specific features in the target and non-target data domains with certain general features that are used frequently in both domains and are relevant to the opinion class. As a result, even if a feature in the target domain has never occurred in the non-target domain, the class label can be predicted by looking up its corresponding feature(s) in the non-target domain.

A study by Tan, Cheng, Wang, and Xu (2009) is the most similar in spirit to this research. It made use of general features in both the target and non-target domains to address the domain adaptation problem in opinion classification. Their approach differed from the study by Blitzer et al. (2007) in that only labeled data in the non-target domain were used with an SSL algorithm, EM-NB, that put more weight on target data for opinion classification. Regardless of their positive contributions to sentiment analysis, both of these domain adaptation strategies involve sophisticated and expensive methods for selecting general features and applying them to sentiment analysis. Believing sentiment is a sentence-level feature, this study conducts opinion classification on the sentence-level, instead of on the document level as in Tan et. al.'s work.

3. EXPERIMENTS

3.1. Selection of Datasets

Three types of text have been explored in prior sentiment analysis studies: news articles, online reviews, and online discourse in blogs or discussion forums. These texts differ from one another in terms of structure, text genre (e.g., level of formality), and the proportion of opinions each contains. A dataset of each type was selected in order to investigate the robustness and adaptability of SSL algorithms for opinion classification and to test the feasibility of SSL for domain adaptation. A small set of blog data was also used for parameter optimization. Several manually created opinion lexicons used in earlier studies were also collected in order to increase classification precision for data domains where opinion detection is particularly difficult.

One of the standard datasets in sentiment analysis is the movie review dataset created by Pang and Lee (2004).[2] It contains 5,000 subjective sentences or snippets from the Rotten Tomatoes[3] pages and 5,000 objective sentences or snippets from IMDB[4] plot summaries, all in lowercase. Sentences containing less than 10 tokens were excluded, and the dataset was labeled automatically by assuming opinion inheritance.

The news article dataset[5] created by Wiebe, Bruce, and O'Hara (1999) is widely used as the gold-standard corpus in opinion detection research. They chose the *Wall Street Journal* portion of the Penn Treebank III (Marcus, Santorini, Marcinkiewicz, & Taylor, 1999) and manually augmented it with opinion related annotations. According to their coding manual, subjective sentences are those expressing evaluations, opinions, emotions, and speculations. For this research, 5,297 subjective sentences and 5,174 objective sentences were selected based on the presence or absence of manually labeled subjective expressions.

The JDPA corpus[6] (Kessler, Eckert, Clark, & Nicolov, 2010), a new opinion corpus released in 2010, consists of blog posts expressing opinions about automobiles and digital cameras. Opinions about named entities (e.g., "seat," "lens") were manually annotated. All sentences containing sentiment-bearing expressions were extracted and objective sentences were manually identified by eliminating subjective sentences that were not targeted to any labeled entities. This process produced 10,000 subjective sentences and 4,348 objective sentences. To balance the number of subjective and objective sentences, 4,348 subjective sentences were randomly selected from the original set of 10,000.

From 2006 through 2008, a dataset called Blogs06[7] was used for tasks in TREC's Blog track. Researchers at the University of Glasgow crawled the blogosphere over an 11-week period from December 2005 to February 2006 to create the Blogs06 collection (Ounis, Rijke, Macdonald, Mishne, & Soboroff, 2007). In this collection, permalink documents (i.e., Web pages containing a single blog post with its associated comments) were the retrieval and assessment units. For TREC's Blog track opinion retrieval tasks, 50 topics (i.e., search queries and descriptions) were released every year, and each participant in the Blog track was to submit several retrieval runs, each run consisting of the top 1000 documents retrieved for each topic. The top documents retrieved across systems for each topic were then manually labeled as topical relevant, topical relevant but not opinion-bearing, and topical relevant and opinion-bearing (i.e., "positive," "negative," or "neutral"). Because topical relevance and opinion polarity would not be taken into consideration in this research, non-relevant data were ignored, and negative, positive, and mixed opinion data were combined into one opinion dataset.

[2] This dataset can be downloaded from http://www.cs.cornell.edu/people/pabo/movie-review-data/, under subjectivity datasets.
[3] http://www.rottentomatoes.com/
[4] http://www.imdb.com/
[5] This dataset can be downloaded from http://www.cs.pitt.edu/mpqa/databaserelease/
[6] The license form for this dataset is available at: http://www.icwsm.org/data/JDPA-Sentiment-Corpus-Licence-ver-2009-12-17.pdf
[7] This dataset can be purchased via this page: http://ir.dcs.gla.ac.uk/test_collections/access_to_data.html

The Blogs06 collection is labeled at the document level and thus required manual labeling to prepare labeled data at the sentence level. In order to avoid bias caused by a particular topic, five TREC labeled opinion-bearing documents (1 positive, 1 negative and 3 mixed opinion) were randomly selected and manually examined for each of the 150 topics, for a total of 750 documents. Because machines cannot be expected to recognize trivial expressions of opinion about which humans are uncertain, emphasis was placed on identifying opinion expressions that contained explicit opinion cues. For example, in a product review, the sentence "I returned this product after a week" may indicate a negative opinion, but it may also state the fact that the product was returned because the reviewer received another as a gift. It is also reasonable to assume that explicit opinion cues may exist around ambiguous opinion expressions to support or explain them (e.g.,"It is horrible! I returned this product after a week."). Therefore, a sentence was labeled as an opinion only if strong traces of opinion cues were present. Sentences that made objective statements were labeled as non-opinion, and the remaining sentences in selected blog posts were ignored. All in all, 1,237 subjective sentences and 616 objective sentences were collected.

3.2. Domain Independent Opinion Lexicons

Several studies have suggested that the use of high-quality opinion lexicons can yield high precision for opinion detection. Therefore, it is advisable to apply these lexicons to boost the classification precision of the initial classifier for SSL runs, especially for difficult data domains such as blog posts. Accordingly, six domain independent opinion lexicons that had proven useful in previous opinion mining studies were collected for use in these experiments.

Adjectives are often connected to the expression of attitudes and have been reported to have a positive and statistically significant correlation with subjectivity (Wiebe et al., 1999). Three adjective opinion lexicons were selected for this research: Index of General Inquirer (IGI) tag categories, a manually constructed list that contains 765 positive and 873 negative words (Stone, 1997); Colin adjectives, an opinion lexicon distributed by Hatzivassiloglou and Wiebe (2000), which include manually and automatically identified semantic oriented adjectives, dynamic adjectives, and gradable adjectives; and strong semantic oriented adjectives in the subjectivity term list created by Wilson, Pierce and Wiebe (2003). Dynamic adjectives were separated from other Colin adjectives into an individual lexicon because of their unique features and their significant contributions.

Appraisal groups have also been suggested as useful in identifying what is called an *appraisal expression*, "a textual unit expressing an evaluative stance towards some target" (Bloom, Garg & Argamon, 2007, p. 308). Given the high cost of full syntactic parsing and the difficulty of fine-level analysis, this research used only the head adjectives, which are marked as positive or negative in the hand-built lexicon distributed by Bloom et al. (2007).

Although not as significant as adjectives, verbs have also been found to be good indicators of opinion information. *Verb classes*, categories for classifying verbs syntactically and/or semantically, are often used for culling opinionated verbs. *Levin's verb classes*, developed on the basis of both intuitive semantic groupings and participation in valence, or polarity alternations (Levin, 1993), are the most popular verb classes used as opinion evidence. For this research, verbs from opinion-related Levin's verb classes, including *judgement* (e.g., "abuse," "acclaim"), *complain* (e.g., "hate," "despise"), and *psych* (e.g., "amuse," "admire," "marvel (at)"), were selected. Similarly, *FrameNet* (Fillmore & Baker, 2001), which groups words, including verbs, according to conceptual structures, provides semantic frames such as *communication* (e.g., "indicate," "convey") as evidence of opinion (Breck, Choi, & Cardie, 2007). For this research, several frames were selected: "agree or refuse to act," "be in agreement on assessment," "desirability," "experiencer (objective /subjective)," "judgment," "opinion," "prevarication," and "statement."

In addition to single words, opinion lexicons used in this research include patterns such as IU collocations (Yang et al., 2007) and bigrams. IU collocations are n-grams with first-person pronouns (e.g., "I," "we") and second-person pronouns (e.g., "you")

as anchor terms. During their experiments for TREC's Blog track, Yang et al. (2007) found that IU collocations worked best as single features. The UMass Amherst Linguistics Sentiment Corpora (Constant, Davis, Potts, & Schwarz, 2009; Potts & Schwarz, 2008) consists of unigrams and bigrams gathered from online book reviews on Amazon[8] and online hotel reviews on TripAdvisor.[9] For each n-gram, total occurrence is reported on an ordinal scale of 1 to 5, with 1 indicating a highly negative review and 5 indicating a highly positive review. In order to pick opinion n-grams, bigrams were excluded if they: contained domain stop words (e.g., book, hotel); occurred frequently at all rating levels; occurred more often at neutral ratings than at either positive or negative ratings; or contained digits or less than 3 characters. Only those n-grams appearing in both Amazon book reviews and TripAdvisor hotel reviews were retained.

Altogether, nine domain-independent opinion lexicons were utilized: appraisal semantic oriented adjectives,[10] gradable and semantic oriented Colin adjectives, dynamic adjectives,[11] IGI semantic oriented adjectives,[12] Wilson subjective terms,[13] Levin's opinion-related verb class terms, FrameNet opinion related category labels, IU collocations, and review bigrams.

3.3. Data Preprocessing

All words in datasets were converted to lower case, and numbers were replaced with the placeholder "#". Unigrams and bigrams were generated for each sentence, and common stop words such as articles and prepositions were removed from unigrams. No stemming was conducted since the literature shows no clear gain from stemming in opinion detection; stemming may actually erase subtle opinion cues such as past tense verbs. For each sentence, nine lexicon scores were assigned, with each score corresponding to the total occurrence of a term in one particular lexicon.

As illustrated in Figure 2, each dataset was randomly split into three portions: 5% of the sentences were reserved as the evaluation set (E) and were available only for S^3VM runs; 90% were treated as unlabeled data (U); and i% (i = 1, 2, 3, 4 or 5) were treated as labeled data (L).

3.4. Experiment Design

In the experiments reported here, opinion detection was treated as a binary classification problem with two categories: subjective sentences (i.e., positive examples, or *p*) and objective sentences (i.e., negative examples, or *n*).

Two groups of experiments were conducted. One group of experiments applied self-training only with the target data domain to investigate the overall feasibility and effectiveness of self-training in opinion detection. The other group of experiments used opinion-labeled data from non-target data domains to examine the applicability of self-training for domain adaptation.

3.4.1 Design of Experiment 1: Basic Self-training

In order to test the effectiveness of self-training with respect to the number of available labeled data, each self-training opinion classifier was trained on i% of the labeled dataset L and the unlabeled dataset U. The corresponding baseline supervised opinion classifier was constructed using only L, and the fully supervised opinion classifier was constructed by treating all data in U and L as labeled

[8] http://www.amazon.com/
[9] http://www.tripadvisor.com/
[10] The appraisal adjectives can be downloaded from http://lingcog.iit.edu/arc/appraisal_lexicon_2007a.tar.gz
[11] The gradable and semantic oriented Colin adjectives and the dynamic adjectives can be downloaded from http://www.cs.pitt.edu/~wiebe/pubs/coling00/coling00adjs.tar.gz
[12] The IGI words can be accessed at http://www.wjh.harvard.edu/~inquirer/inqdict.txt. Positive and negative words were extracted.
[13] The Wilson subjective terms are included in the OpinionFinder package available at http://www.cs.pitt.edu/mpqa/opinionfinderrelease/. Strong subjective terms were extracted.

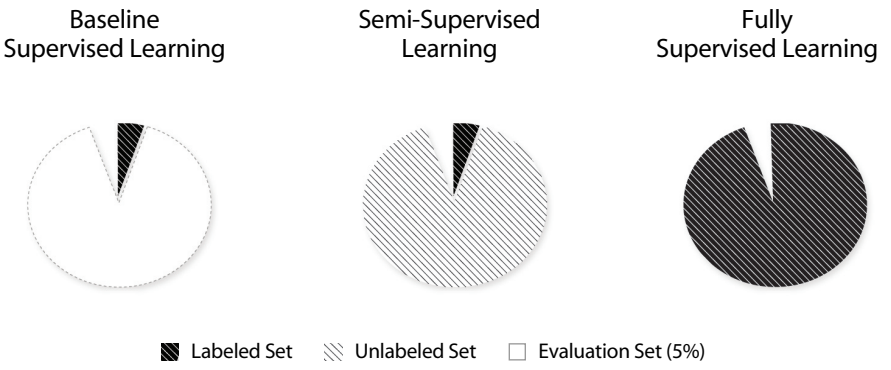

Fig. 2 Data Split for Semi-supervised Learning Runs, Baseline Supervised Learning Runs and Fully Supervised Learning Runs

data. Performance of each self-training run was compared with the performance of both the baseline SL run and the full SL run.

Although both SVM and Naïve Bayes algorithms are widely used for document classification, the Naïve Bayes classifier was selected as the base classifier for this study because preliminary experiments showed that, even with a logistic model to output probability scores for the SVM classifier, the difference in probabilities is too small to select a small number of top classification predictions. The multinomial Naïve Bayes classifier in Weka (Hall, Frank, Holmes, Pfahringer, Reutemann & Witten, 2009) was used to run all the experiments.

For each sentence, both unigrams and bigrams were extracted as classification features. Higher order n-grams (i.e., n>=3) were not used because effective high order n-grams cannot be extracted from a small labeled dataset. Binary values (i.e., presence or absence) were applied for these features.

Other parameter settings included: (1) for all self-training runs, iterations stopped when there were no more unlabeled data; (2) for each iteration, a number of unlabeled examples u, smaller than U, were randomly extracted from the unlabeled dataset U for classifiers to predict opinion labels; and (3) for each iteration, opinion examples (p) and non-opinion examples (n) were added back to the labeled dataset. The ratio between p and n approximates the distribution of opinions and non-opinions in the labeled dataset. The optimized parameter values established in this group of experiments will be used for the next group of experiments.

3.4.2 Design of Experiment 2: Domain Adaptation

Because movie review data are often labeled with sentiment classes and are reported to achieve great classification accuracy in the sentiment analysis literature, they were treated as the source data, while datasets for news articles and blog posts were treated as target data.

While the data split for the target domain was the same as that used in previous experiments, all sentences in the source domain, except for the 5% evaluation data, were treated as labeled data. For example, in order to identify opinion-bearing sentences from the blog dataset, all 9,500 movie review sentences and i% of blog sentences were used as labeled data, 90% of blog sentences were used as unlabeled data, and 5% were reserved as evaluation data. In addition, a parameter was added to gradually reduce the weight of non-blog examples in the training set during iterations, similar to the approach taken by Tan et al. (2009). To reduce bias caused by features specific to one non-target data domain, labeled data from two different non-target data domains were combined as training data for both supervised and semi-supervised learning algorithms (i.e., in co-training, two view classifiers were trained on two non-target domains).

In order to compare the benefits of employing

non-target labeled data to the benefits of using general opinion lexicons to deal with the domain transfer problem, another set of domain adaptation experiments used general opinion lexicons instead of borrowing opinion labeled sentences from other domains. In addition to the n-gram features, SL and SSL runs in this set used features from nine opinion lexicons to represent each in-domain sentence.

3.5. Evaluation Measures

Classification accuracy was used as the evaluation measure when comparing SSL and SL runs. Classification accuracy evaluates the overall correctness of a classifier and is calculated using the formula ACC = (a+d)/(a+b+c+d).

In addition, two measures were adopted to determine whether performance increased when more unlabeled data were used and whether the contribution of unlabeled data decreased with the increase in available labeled data, as suggested in most SSL studies (e.g., Nigam & Ghani, 2000).

4. RESULTS AND DISCUSSION

4.1. Preliminary Experiments

Self-training runs with various parameter settings were conducted on TREC's blog data to evaluate the impact of different experimental settings and to determine optimized parameters for all self-training runs.

4.1.1. Feature Selection

Two popular feature selection methods—information gain (IG) and chi-square (CHI)—were investigated. When keeping all other parameters fixed and selecting the top 100 features, neither feature selection method contributes to SSL performance with labeled data from 1% to 5% of the total dataset. Because feature selection consumes computing time, especially when a new classification model must be built for each iteration, no feature selection was conducted for the subsequent experiments.

4.1.2. Unlabeled Data Available for Each Iteration

To decide how many unlabeled sentences u should be available to the classifier on each iteration, experiments were designed using 20, 75, 100 and all unlabeled sentences (i.e., approximately 1660 sentences). By computing the average improvement of self-training runs over corresponding baseline SL runs with 1% to 5% labeled data, it was found that self-training runs classifying all unlabeled sentences on each iteration decreased classification accuracy by 4.67%; self-training runs classifying 100 unlabeled sentences on each iteration increased baseline performance by 2.08%; self-training runs classifying 75 unlabeled sentences on each iteration did not improve baseline performance; and self-training runs classifying only 20 unlabeled sentences on each iteration increased baseline performance by 4.18%. For the following experiments, u was set to 20.

After p auto-labeled opinion sentences and n auto-labeled non-opinion sentences were selected and added to the labeled dataset, p+n unlabeled sentences can be drawn from U to replenish u or a new set of u can be generated from U. Experiments using TREC's blog dataset indicated that replenishing u outperformed generating a new set of u by 11.87% in terms of classification accuracy. One explanation is that, for succeeding iterations, replenishing u kept those unlabeled sentences for which the classifier generated low prediction scores in the current iteration and forced the classifier to reclassify difficult sentences, while generating a new set of u allowed the classifier to select sentences that were easy to classify.

On the one hand, in order to avoid mislabeled data in the labeled dataset, only the most confidently labeled data should be selected, and a small value for p and n would be preferred. Alternatively, in order to reduce the number of iterations necessary for SSL to converge, a larger value for p and n would be preferred. Preliminary experiments compared the results of setting p and n either to one or to two and found no noticeable difference. For this reason, p and n were set at two for all experiments.

4.2. Basic Self-training

The first experiment examined the effectiveness of self-training that used only in-domain data. For the movie review, news, and blog data domains, the

performance of self-training was compared with the performance of SL runs, which used the same number of labeled sentences as well as those that used all data as labeled sentences.

Table 1 shows the classification accuracy of self-training and two supervised learning runs for movie reviews. The more labeled data provided for the baseline SL runs, the better the performance: With 100 labeled sentences, the baseline SL run achieved classification accuracy of only 63.80%; but with 500 labeled sentences, the supervised learning classifier achieved classification accuracy of 80.20%. The second row shows the performance of the simple self-training method using 100 to 500 labeled sentences and an additional 9000 unlabeled sentences. These self-training runs improved performance over the corresponding baseline supervised runs: For example, using 100 labeled sentences, self-training achieved a classification accuracy of 85.2% and outperformed the baseline SL by 33.5%. Although the full SL run using all labeled data surpassed the simple self-training run by 4.9%, significant effort was saved by labeling only 100 sentences rather than 9,100. If approximately 30 seconds are needed to label each sentence, self-training saves 225 hours of labor of three human annotators; and if they are paid $15/hour, this saves almost $3,400. With 500 labeled sentences, self-training improved accuracy over the baseline supervised run by 6%, indicating that self-training is particularly beneficial when the number of labeled data is small.

Overall, self-training which iteratively labeled unlabeled data with one classifier was effective for movie reviews and achieved performance close to fully supervised learning while saving the labor involved in labeling thousands of unlabeled sentences. Because news articles follow similar patterns, their results will not be shown here.

As shown in Table 2, none of the self-training runs proved beneficial in the blog domain. This is because the blog data domain is even more challenging than the news domain. The language used in blog posts is more informal than the language of the other two data domains, and blog writing contains a variety of opinion cues not found in movie reviews or news writing. Furthermore, because the JDPA blog data are focused on reviews of cars and cameras, opinion and non-opinion sentences share topic-related features; moreover, the average length for opinion and non-opinion sentences in blog posts is 17 words, shorter than that for movie reviews (23.5 words) or news articles (22.5 words). In fact, approximately one quarter of the sentences in the blog dataset had only 5 to 10 words. This poses an additional challenge because there is less information for the classifier in terms of the number of individual features.

With limited labeled data, the results of these experiments suggest that self-training can make effective use of unlabeled data for opinion detection in certain data domains (e.g., movie reviews) but not in others (e.g., news and blog data). One reason for the failure of self-training in the blog domains is the low classification accuracy of initial runs: The

Table 1. Classification Accuracy (%) of Self-training and SL Runs for Movie Reviews

Run Type	# of Original Labeled Sentences				
	100	200	300	400	500
Baseline SL	63.80	73.60	77.20	79.40	80.20
Self-training	85.20	86.60	87.00	87.20	85.20
Full SL	90.00	92.00	91.80	91.60	91.80

Note. Settings for self-training: u=20, p=2, n=2, n-grams=unigrams+bigrams.
Full SL runs used an additional 9000 labeled sentences.

Table 2. Classification Accuracy (%) of Self-training and SL Runs for Blog Posts

Run Type	# of Original Labeled Sentences				
	86	**172**	**258**	**344**	**430**
Baseline SL	55.05	58.95	61.93	64.69	66.06
Self-training	54.59	55.73	56.65	58.49	64.45
Full SL	71.56	73.17	72.71	72.94	72.48

Note. Settings for self-training: *u*=20, *p*=2, *n*=2, n-grams=unigrams+bigrams. Full SL runs used an additional 7740 labeled sentences.

performance of blog baseline classifiers was only slightly better than chance (50%) and decreased the quality of auto-labeled data.

4.3. Domain Adaptation

In order to deal with challenging data domains such as blog posts, one possible solution is to improve baseline accuracy for self-training by introducing high-quality features: for example, augmenting the feature set with domain independent opinion lexicons such as those which have been suggested as effective in creating high precision opinion classifiers. An alternative approach for dealing with challenging data domains is to borrow labeled data from one or more "easy" domains: for example, the use of movie review data in self-training applications for opinion detection in news article and blog domains.

4.3.1. Using Domain-Independent Opinion Lexicons

In addition to unigram and bigram features with binary values, nine lexicon features were added to the feature set. To avoid the possibility that the large number of n-gram features would weaken these nine lexicon features, the value of each lexicon feature (e.g., dynamic adjectives) was not binary but represented the total number of matches between lexicon terms and the words in a target sentence. For example, the value of Wilson lexicon features for the sentence "I like these two much better than the versions made for the Hong Kong market" is two because two Wilson lexicon terms, 'like' and 'better,'are used in this sentence. Redundancy between lexicons was not removed under the assumption that one word occurring in multiple lexicons makes it a strong opinion indicator. For example, 'like' is included in the Levin verb class lexicon, the frameNet lexicon, and the Wilson lexicons, and its occurrences were counted when calculating values for all three lexicon features.

In Table 3 and Table 4, the baseline supervised learning runs using domain-independent opinion lexicon features (i.e., Baseline SL w/ Lexicon) produced higher classification accuracies than baseline supervised learning runs that did not use lexicon features (i.e., Baseline SL w/o Lexicon). However, self-training runs that used opinion lexicons (i.e., Self-training w/ Lexicon) did not generally improve the baseline run (i.e., Baseline SL w/ Lexicon); in some cases, performance was even lower than that of the corresponding self-training runs that did not use domain-independent opinion lexicon information (i.e., Self-training w/o Lexicon). For example, using opinion lexicon features with 86 labeled blog sentences, supervised learning yielded a classification accuracy of 63.76%, 8.71% higher in absolute value than the classification accuracy produced by the supervised learning run that made no use of opinion lexicon features; however, after self-training iterations, the performance of the former run decreased to 51.38%, 3.21% lower in the absolute value of classification accuracy than the classification accuracy produced by the latter run. This may be because, as a closer look at the distribution of opinion lexicon terms in the three datasets indicates, many opinion lexicon terms actually occur frequently in objective, non-opinion sentences.

Table 5 shows the number of unique opinion lexicon terms that appear in subjective and objective

Table 3. Classification Accuracy (%) of Self-training With and Without Opinion Lexicon Features for News Articles

Run Type	# of Original Labeled Sentences				
	103	206	309	412	515
Baseline SL w/o Lexicon	60.50	64.31	69.47	69.47	71.38
Self-training w/o Lexicon	60.11	65.84	66.41	67.75	67.56
Baseline SL w/ Lexicon	66.60	70.42	70.99	72.14	72.52
Self-training w/ Lexicon	59.73	66.41	71.18	70.61	70.61

Note. Settings for self-training: *u*=20, *p*=2, *n*=2, n-grams=unigrams+bigrams.

Table 4. Classification Accuracy (%) of Self-training With and Without Opinion Lexicon Features for Blog Posts

Run Type	# of Original Labeled Sentences				
	86	172	258	344	430
Baseline SL w/o Lexicon	55.05	58.95	61.93	64.69	66.06
Self-training w/o Lexicon	54.59	55.73	56.65	58.49	64.45
Baseline SL w/ Lexicon	63.76	64.68	63.53	66.51	67.89
Self-training w/ Lexicon	51.38	62.16	55.73	61.47	69.04

Note. Settings for self-training: *u*=20, *p*=2, *n*=2, n-grams=unigrams+bigrams.

data in the three data domains as well as the total occurrence of opinion lexicon terms in subjective and objective sentences. Although opinion lexicon terms are used more often in opinion sentences than in non-opinion sentences, their presence does not appear to be a strong indicator of opinions. For example, more than half of the opinion lexicon features that appear in opinion blog sentences also appear in non-opinion blog sentences. When considering their total occurrence, opinion lexicon terms are used in opinion sentences approximately three times as often as in non-opinion sentences in both the blog and news data domains; opinion lexicon terms are used in non-opinion sentences a little more than half as often as they are in opinion sentences in the movie review domain. This suggests that automatically created subjective and objective movie review data will not necessarily reflect opinion and non-opinion classes.

Table 5. Distribution of Domain Independent Opinion Lexicons

# of Matches	Dataset					
	Movie Reviews		News Articles		Blog Posts	
	Non-Op	Op	Non-Op	Op	Non-Op	Op
Unique Terms	1076	1428	502	1127	459	753
Total Occurrence	4867	8596	1865	5576	1778	4668

Note. Non-Op=non-opinion; Op=opinion.

The inefficiency of opinion lexicons can be attributed to the fact that opinion features are often very sensitive to the context in which they occur. For example, "like" is included in three opinion lexicons and is therefore treated as a good opinion indicator, but when it is used in the sentence "the lens cap finally snaps into the front of the lens like other makers' models," it is no longer an opinion indicator. As a result, when there was a limited number of labeled data at the beginning of a self-training run, extra opinion lexicon features helped; however, with more and more unlabeled data labeled automatically and used to replenish the labeled dataset, the limitations of opinion lexicons were amplified, undermining overall performance.

4.3.2. Using Labeled Data in Non-Target Domain

A preliminary experiment on the use of movie review data was conducted on the news domain. This analysis was followed by a more in-depth investigation of the use of movie review data in the blog data domain.

From Movie Reviews to News Articles

This experiment tested an extreme situation where there were no labeled data available in the target data domain. To begin, 9,500 labeled movie review sentences were used to train a Naïve Bayes classifier. Although this classifier produced a fairly good classification accuracy of 89.2% on movie review data, its accuracy in a domain-transfer SL run on news data was poor (64.1%), demonstrating the severity of the domain transfer problem.

A self-training run starting with the same Naïve Bayes classifier trained on movie review data and using unlabeled data from the news domain (i.e., a domain-transfer SSL run) showed some improvement, achieving a classification accuracy of 75.1% that surpassed the domain-transfer SL run by more than 17% with no extra efforts for manual annotation. To further understand how well SSL handles the domain transfer problem, a full SL run that used all labeled news sentences was also performed. This full SL run achieved 76.9% classification accuracy, only 1.8% higher in absolute value than the domain-transfer SSL run, which had not used any labeled news data.

From Movie Reviews/News to Blog Posts

Domain transfer self-training runs for blog data combined all movie review data and i% labeled blog data to form the initial labeled dataset, and then followed the traditional self-training procedure. A control factor was introduced and investigated to gradually reduce the impact of out-of-domain data (i.e., movie reviews) on each iteration.

Table 6 reports the results of self-training runs to identify opinion sentences in blog posts, both with and without the use of movie review data, as well as corresponding baseline and fully supervised learning runs. The results for baseline SL runs without movie reviews and self-training without movie reviews show that self-training using only blog data decreases baseline SL performance. By keeping the same settings and adding more labeled data from the movie review domain, self-training with movie reviews increased the performance of SL runs by 12% to 15% and came closer to the performance of full SL runs, which used 90% of the labeled blog data. In the case of domain transfer runs, the number of available in-domain labeled data did not appear to have an impact on overall performance: neither supervised nor semi-supervised runs using movie review data produced higher classification accuracies with increasing numbers of labeled blog sentences. For example, the self-training run using movie review data yielded the same classification accuracy of 71.10% with as few as 86 or as many as 430 labeled blog sentences in the original training set. This may be due to the preponderance of movie review data available during training.

A control factor intended to reduce the bias of movie review data was added to weaken the effects of domain transfer gradually (i.e., a decrease of 0.001 on each iteration). The results reported for self-training runs with both movie review data and weight control show that these runs outperformed runs that did not use weight control by 1% to 3%, reaching and occasionally exceeding the performance of the full SL run.

Overall, for high-challenge data domains, adoption of domain independent opinion lexicons resulted in only minimal improvement, but applying simple self-training alone was promising for tackling domain transfer from the source domain of

Table 6. Classification Accuracy (%) of Self-training With and Without Labeled Movie Reviews

Run Type	# of Original Labeled Sentences (Blog)				
	86	172	258	344	430
Baseline SL	55.05	58.95	61.93	64.69	66.06
Self-training	54.59	55.73	56.65	58.49	64.45
Baseline SL w/ m.r.	63.07	62.16	62.61	62.16	61.70
Self-training w/m.r.	71.10	70.87	71.41	70.41	71.10
Self-training w/m.r. w/w.c.	72.94	72.94	72.48	71.56	71.79
Full SL	71.56	73.17	72.71	72.94	72.48

Note. m.r. = Movie reviews. w.c. = Weight Control.
Settings for self-training: u=20, p=2, n=2, n-grams=unigrams+bigrams.
Full SL used an additional 7740 labeled blog sentences.

movie reviews to the target domains of news articles and blog posts. Supported by the opinion feature distribution statistics in Table 5, one guess for the success of movie reviews in helping classifying opinions in news articles and blogs is the rich opinion features in this data domain.

5. CONCLUSION

Sentiment is an important aspect of many types of information and being able to identify and organize sentiments is essential for information studies. The shortage of labeled data has become a severe challenge for developing effective sentiment analysis systems. This study tackled this challenge by investigating a semi-supervised learning (SSL) approach, motivated by limited labeled data and the availability of plentiful unlabeled data. Specifically, this research investigated self-training strategies in dealing with the domain transfer problem via learning unlabeled data in the target domain and labeled data in non-target domain(s).

To understand the feasibility and effectiveness of SSL for sentiment analysis, self-training was applied to three datasets from domains with different characteristics (i.e., movie reviews, news articles, and blog posts), and its performance varied across domains. For movie reviews, all self-training runs showed the advantage of using unlabeled data for opinion detection with both time and cost benefits. Due to the nature of the movie review data, opinion detection in movie reviews is an "easy" problem because it involves genre classification and thus relies, strictly speaking, on distinguishing movie reviews from plot summaries. For other manually created datasets that are expected to reflect real sentiment characteristics, self-training was impeded by low baseline precision and demonstrated only limited improvement. Blog posts are the most challenging domain and blog data showed no benefits from implementing self-training. However, with the addition of out-of-domain labeled data (i.e., movie review data), self-training for identifying opinion sentences in blogs exceeded fully supervised learning using all available labeled blog data. This promising result suggests great value in further exploration of SSL for domain adaptation, especially because of its easy implementation.

The contributions of this research are four-fold. First, the findings of this research indicate a general approach that can be adapted for use in existing sentiment analysis systems across data domains and across languages. These findings also provide valuable guidelines and evaluation baselines for later studies applying SSL algorithms in sentiment analysis. Second, there are several applications for automatically labeled data generated by the effective SSL strategies reported in this research: creating sentiment labeled corpora directly; providing candidates

for manual annotation; and extracting sentiment-bearing features. Third, the SSL strategies investigated in this research, especially those related to domain adaptation, are readily extensible to other text mining systems (e.g., genre identification). Finally, this research contributes to SSL re-search by expanding the spectrum of SSL applications to include sentiment analysis, confirming the effectiveness of SSL as a general approach for dealing with insufficient quantities of labeled data, and providing promising new approaches for domain adaptation.

REFERENCES

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. *ACM Transactions on Information Systems, 26*(3).

Aue, A., & Gamon, M. (2005). Customizing sentiment classifiers to new domains: A case study. In *Proceedings of International Conference Recent Advances in Natural Language Processing (RANLP-2005)*, Borovets, Bulgaria, 21-23 September 2005.

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (pp. 440-447). Association for Computational Linguistics.

Bloom, K., Garg, N., & Argamon, S. (2007). Extracting appraisal expressions. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT)*, Rochester, NY (pp. 308-315). Morristown, NJ: Association for Computational Linguistics.

Bollen, J., Mao, H., & Zeng, X. J. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, *2*(1), 1-8.

Breck, E., Choi, Y., & Cardie, C. (2007). Identifying expressions of opinion in context. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, Hyderabad, India, 6-12 January 2007 (pp. 2683-2688).

Chesley, P., Vincent, B., Xu, L., & Srihari, R. K. (2006). Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches to Analyzing Weblogs*, Stanford University, CA., 27-29 March 2006, Menlo Park, CA: AAAI Press.

Conrad, J. G., & Schilder, F. (2007). Opinion mining in legal blogs. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, Stanford, CA (pp. 231-236). New York, NY: ACM.

Constant, N., Davis, C., Potts, C., & Schwarz, F. (2009). The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung 33*, 5-21.

Fillmore, C. J., & Baker, C. F. (2001). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh, PA.

Gamon, M. (2004). Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland, 23-27 August 2004. Stroudsburg, PA, USA: Association for Computational Linguistics.

Hatzivassiloglou, V., & Wiebe, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th Conference on Computational Linguistics*, Saarbrücken, Germany, 31 July-4 August 2000 (pp. 299-305). Stroudsburg, PA, USA: Association for Computational Linguistics.

He, Y., & Zhou, D. (2011). Self-training from labeled features for sentiment analysis. *Information Processing and Management*, *47*(4), 606-616.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. ACM *SIGKDD Explorations news letter*, *11*(1), 10-18.

Kessler, J. S., Eckert, M., Clark, L., & Nicolov, N. (2010). The ICWSM 2010 JDPA sentiment corpus for the automotive domain. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media Data Workshop Challenge (ICWSM-DWC)*, Washington, D.C., USA.

Ku, L. W., & Chen, H. H. (2007). Mining opinions from the Web: Beyond relevance retrieval.

Journal of the American Society for Information Science and Technology, *58*(12), 1838-1850.

Levin, B. (1993). *English verb classes and alternations*. Chicago, IL: University of Chicago Press.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., & Taylor, A. (1999). *Treebank-3*. Linguistic Data Consortium, Philadelphia.

Nigam, K., & Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of the Ninth International Conference on Information and Knowledge Management* (pp. 86-93). New York, NY, USA: ACM.

Ounis, I., Macdonald, C., & Soboroff, I. (2008). Overview of the TREC-2008 Blog Track. In *Proceedings of the 17th Text REtrieval Conference (TREC 2008)*.

Ounis, I., Rijke, M. D., Macdonald, C., Mishne, G., & Soboroff, I. (2007). Overview of the TREC-2006 Blog track. In *Proceedings of the 15th Text REtrieval Conference*.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Barcelona, Spain, 21-26 July 2004, (pp. 271-278). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, Philadelphia*, PA, 6-7 July 2002, (pp. 79-86). Stroudsburg, PA, USA: Association for Computational Linguistics.

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. New York, NY: Bloomsbury Press.

Potts, C., & Schwarz, F. (2008). *Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora*. Ms.: UMass Amherst.

Riloff, E., & Jones, R. (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, Orlando, FL (pp. 474-479). Menlo Park, CA, USA: American Association for Artificial Intelligence.

Stone, P. J. (1997). Thematic text analysis: New agendas for analyzing text content. In C. Roberts (Ed.), *Text analysis for the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Tan, S., Cheng, X., Wang, Y., & Xu, H. (2009). Adapting Naïve Bayes to domain adaptation for sentiment analysis. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, (pp. 337-349).

Tsou, B. K. Y., Yuen, R. W. M., Kwong, O. Y., Lai, T. B. Y., & Wong, W. L. (2005). Polarity classification of celebrity coverage in the Chinese press. In *Proceedings of the International Conference on Intelligence Analysis*, McLean, VA, 2-4 May 2005.

Wiebe, J., Bruce, R., & O'Hara, T. P. (1999). Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, College Park, MD, 20-26 June 1999 (pp. 246-253). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the 6th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2005)*, Mexico City, Mexico, 13-19 February 2005 (pp. 486-497). Heidelberg, Berlin: Springer-Verlag.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics*, *30*(3), 277-308.

Wilson, T., Pierce, D. R., & Wiebe, J. (2003). Identifying opinionated sentences. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Demonstrations*, Edmonton, Canada (pp. 33-34). Stroudsburg, PA, USA: Association for Computational Linguistics.

Yang, K., Yu, N., & Zhang, H. (2007). WIDIT in TREC-2007 Blog track: Combining lexicon-based methods to detect opinionated blogs. In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*.

Yu, N., & Kübler, S. (2010). Semi-supervised learning for opinion detection. In *Proceeding of the IEEE/ WIC/ ACM International Conference on Web*

Intelligence and Intelligent Agent Technology, vol. 3, Toronto, ON, Canada, 31 August – 3 September 2010 (pp. 249-252). Stroudsburg, PA, USA: Association for Computa-tional Linguistics.

Yu, N., & Kübler, S. (2011). Filling the gap: Semi-supervised learning for opinion detection across domains. In *Proceeding of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland, OR, 23-24 June 2011 (pp. 200-209).

Yu, N., Kübler, S., Herring, J., Hsu, Y. Y., Israel, R., & Smiley, C. (2012). LASSA: Emotion detection via information fusion. *Biomedical Informatics Insights, 5*(Suppl. 1), 71-76.

Zhang, W., & Yu, C. (2007). UIC at TREC 2007 Blog track. In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*.

Zhu, X. (2008). *Semi-supervised learning literature survey*. Department of Computer Sciences, University of Wisconsin, Madison. (Technical Report No. 1530).