ASPECT-BASED OPINION MINING OF CUSTOMER REVIEWS

ZHEN HAI

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfilment of the requirement for the degree of
Doctor of Philosophy

2014

Acknowledgments

I would like to express my deep gratitude to my supervisors Dr. Gao Cong and Dr. Kuiyu Chang. This thesis would not have been possible without their guidance, encouragement, and support. I would like to thank Dr. Qinbao Song, Dr. Jung-jae Kim, and Dr. Christopher C. Yang for their guidance and constructive suggestions on my work. I also thank the thesis committee for their insightful comments and suggestions.

I am fortunate to work with a number of superb colleagues in our lab, and I would like to express my thanks to Peilin Zhao, Wenting Liu, Guangxia Li, Shaohua Li, Quan Yuan, Zhiqiang Xu, and Chang Xu. I would also like to express my thanks to my friends Yuanyuan Guo, Guangtao Wang, Chundong Wang, Jerry Caozhen Zhang, Jun Gu, Ning Chen, Xin Cao, Xingpeng Xu, Shijie Xiao, Xiaojun Yu, Lisi Chen, Fan Zhang, Yao Zhang, and Feng Li for their friendship and help.

I thank Nanyang Technological University (NTU) for offering me the NTU Research Scholarship. My comfortable life in NTU would not have been possible without the support of the scholarship.

I would like to express my deepest gratitude to my family, whose support and care have encouraged me to overcome the difficulties to complete my doctoral study. Finally, I would like to thank Ms. Peng Cheng. Her kindness, patience, encouragement, and care sustain me.

Contents

Abstract

Online reviews are immensely valuable for customers to make informed decisions on product purchase, hotel booking, etc., and for businesses to improve the quality of their products and services. However, customer reviews grow very rapidly in quantity, while varying largely in quality. It is practically impossible for users to read through all the reviews for good decision-making.

Opinion mining, also known as sentiment analysis, has been employed to automatically discover and summarize online reviews. In this thesis, we focus on the problem of aspect-based opinion mining of customer reviews. Our goal is to study and develop computational opinion mining techniques to support users to digest the huge amount of review data. In particular, we study three closely related problems as described below.

The first problem deals with extracting aspect terms and opinion words that appear in customer reviews. We propose a generalized corpus statistics association based bootstrapping approach (ABOOT). ABOOT starts with a small list of annotated aspect seeds, and then iteratively extracts a large number of domain-specific aspect terms and opinion words from a given review corpus. ABOOT is able to work properly with only one seed, which can be simply domain word, e.g., "hotel" for hotel reviews.

Our second problem focuses on identifying implicit aspects for the opinion words devoid of explicit aspects. Implicit aspects refer to the aspects that do not appear but are implied by opinion words in reviews. In opinion mining, very little work has been done on this problem. We propose a cooccurrence association rule mining method (coARM). coARM first discovers a significant set of association rules from a review corpus, and then it applies the rules to the opinion words devoid of explicit aspects for implicit aspect identification.

The third part of this thesis deals with modeling customer reviews with aims at identifying semantic aspects and opinions as well as predicting overall review ratings in a

unified framework. We introduce a new supervised joint topic model named supervised joint aspect and opinion model (SJAOM). SJAOM incorporates the overall ratings as supervision data, and simultaneously models the pairwise aspect terms and opinion words in each review. One key advantage of SJAOM is its ability to jointly identify the semantic aspects and opinions that are predictive of the overall ratings of reviews.

Experimental results on real-world customer reviews demonstrate the benefits of our proposed methods for opinion mining problems, notably the SJAOM model.

Chapter 1

Introduction

We first describe the background and motivation of our research. Then, we present the problems of aspect-based opinion mining as well as the methods proposed for the problems. We conclude this chapter with a summary of our contributions, followed by the organization of this thesis.

1.1 Background and Motivations

Nowadays, with the increasing popularity of e-commerce, more and more people are drawn to online shopping. They are also willing to write reviews to share their hands-on experiences and opinions on the purchased products or services on the e-commerce websites, such as Amazon [1], TripAdvisor [2], and Yelp [3].

Online reviews are very valuable, this is because: 1) They have become an inevitable part of consumers' decision making process on product purchase, hotel booking, etc.; According to a survey, a massive 88% of respondents agreed that they "sometimes or always" consult reviews prior to making a purchase; 2) They collectively form a low-cost and efficient feedback channel for businesses to keep track of their reputations and

[1] www.amazon.com
[2] www.tripadvisor.com
[3] www.yelp.com

customer sentiments, which can be used to improve the quality of their products and services.

However, customer reviews are constantly growing in quantity, while varying largely in quality. A product can easily receive hundreds or even thousands of reviews within a short period of time. Among the large number of customer reviews, high-quality reviews are usually intermixed with the useless ones. It could be very difficult, if not impossible, to manually read through all the reviews for informed decision making.

Opinion mining, also known as sentiment analysis or opinion analysis, is a field of study that analyzes people's opinions, sentiments, comments, attitudes, and evaluations towards entities (e.g., product) [Liu12]. It has been employed to automatically discover and summarize online user-generated reviews.

Opinions or sentiments simply refer to the positive and negative semantic orientations or the degree of praise and criticism demonstrated towards entities. In general, opinions and sentiments expressed in reviews can be analyzed at various levels of granularity. Document-level opinion analysis focuses on classifying the overall sentiment orientation expressed in a whole review document. Similarly, sentence-level opinion analysis aims at analyzing the subjectivity or sentiment expressed in each review sentence. Opinion analysis at phrase/word level mainly focuses on inferring the opinion polarities of individual phrases/words.

As a matter of fact, opinion analysis at document, sentence, or phrase/word level does not discover what exactly people like and dislike in customer reviews. In other words, it cannot associate the identified opinion orientations to the corresponding targets or aspects commented on in reviews. Aspect refers to the unique facet of a reviewed entity (product), and it can be the semantic attribute, component, function, or feature of the entity, e.g., screen of a cellphone. Clearly, the identified opinions and sentiments without their corresponding aspects are of limited use in reality [Liu12].

For example, "The screen is large, and the exterior is also very beautiful, I will recommend it though a little bit expensive!", this review expresses conflicting opinions towards different aspects of a cellphone, though it shows an overall positive opinion. The opinion polarities on the aspects "screen" and "exterior" of this cellphone are positive, which are indicated by the opinion words "large" and "beautiful", while the opinion polarity on the aspect "price" (implicit aspect) is negative, which is indicated by the opinion word "expensive".

Such fine-grained opinion analysis may very well tip the balance in purchase decisions. Nowadays savvy consumers are no longer satisfied with just knowing the overall opinion rating about a product. They want to understand why it receives the particular rating, that is, which positive or negative aspects contribute to the final rating. It is thus significant to identify the specific aspects commented on in reviews and the aspect-specific opinions in practice.

In this thesis, we focus on the problem of aspect-based opinion mining of customer reviews. Our goal is to develop effective computational opinion mining techniques to support users for digesting the huge amount of review data. To this end, we study three closely related problems: 1) Extracting aspect terms and opinion words that appear in reviews via a corpus statistic association based bootstrapping approach (ABOOT), 2) Identifying implicit aspects for the opinion words devoid of explicit aspects via a cooccurrence association rule mining approach (coARM), and 3) Modeling reviews with aims at identifying semantic aspects and opinions as well as predicting overall review ratings via a supervised joint topic model.

1.2 Problems and Methodologies

In this section, we present briefly the three problems to be addressed in this thesis as well as the methods developed for them.

The first problem deals with extracting aspect terms and opinion words that appear in customer reviews. An aspect term refers to the specific term used to express an aspect and typically occurs as noun or noun phrase in online reviews. For instance, the nouns "screen", "LCD", and "display" are particular aspect terms frequently mentioned in cellphone reviews, and they all indicate the single semantic aspect "screen". Opinion words indicate the words used to express opinion and sentiment orientations on aspects, and they typically occur as adjectives in reviews. For example, in this cellphone review "The exterior is so beautiful!", the adjective "beautiful" is an opinion word, which expresses positive opinion on the aspect term "exterior". Extracting aspect terms and opinion words from customer reviews is essential for fine-grained opinion analysis of the reviews.

Semantic dependency relations naturally exist between aspect terms and opinion words, even among aspect terms or opinion words themselves. Syntactic parsing rule based linguistic methods may be used to recognize the pairwise semantic dependency relations to locate aspect terms and opinion words from reviews. However, the methods tend to suffer from: 1) the poor coverage of the manually developed syntactic rules, and 2) informal language of customer reviews, which typically contain misspellings and grammatically incorrect sentences.

We observe that the semantic dependency between each pair of words can be measured, and the results can be exploited to discriminate valid aspect terms and opinion words from invalid ones. We propose to employ corpus statistics association models to quantify the pairwise semantic dependency, and thus introduce a generalized association-based bootstrapping method (ABOOT) for the problem. ABOOT starts with a small set of annotated aspect seeds, then iteratively identifies a large number of aspect terms and opinion words relying on corpus statistics association analysis. Two particular statistical association models, likelihood ratio tests (LRT) [Dun93] and latent semantic analysis (LSA) [DDH90] have been evaluated to compute the pairwise word dependency, which thus lead to two instances of the proposed ABOOT method, namely, LRTBOOT and LSABOOT.

Our second problem focuses on identifying implicit aspects for opinion words devoid of explicit aspect terms in reviews. People often express their opinions but do not explicitly mention the corresponding aspects in reviews. For example, “It is really beautiful.”, the opinion word “beautiful” actually implies an aspect “appearance” of the cellphone, which does not appear explicitly in the sentence. Implicit aspects indicates the aspects that are not explicitly mentioned but are implied by opinion words in review sentences.

Very little research has been done on implicit aspect identification. We propose a two-phase cooccurrence association rule mining approach (coARM) to address this problem. In particular, in the first phase of rule generation, we discover a set of non-trivial cooccurrence association rules in the form of [opinion-word $\rightarrow$ aspect-term] based on a cooccurrence matrix. We then cluster the rule consequents, i.e., aspect terms, to create much robust association rules for each rule antecedent, i.e., opinion word. In the second phase of rule application, for each new recognized opinion word devoid of explicit aspect, we find the best matched association rule, whose rule consequent corresponds to the majority aspect cluster. The representative aspect term of the cluster is accordingly identified as the implicit aspect of the opinion word.

The third problem of this thesis deals with modeling customer reviews with aims at identifying semantic aspects and opinions on the aspects as well as predicting overall review ratings in a unified framework. Recently, probabilistic topic models have been shown effective for discovering semantic topical structure of textual data. Different extensions on a basic latent Dirichlet allocation model (LDA) [BNJ03] have been developed to address opinion analysis problem [TM08b, LH09, ME11]. As far as we know, almost all existing topic modeling approaches are unsupervised.

Generally, a customer review often comes with an overall rating, for instance, in the form of one to five stars. We propose to exploit the overall rating information as supervision knowledge to guide the process of detecting the semantic aspects and aspect-based opinions from reviews. Most of existing topic models for opinion analysis use the bag-of-words representation of a review document. Differently, we reduce each review as a bag of opinion pairs, where each opinion pair consists of an aspect term and its corresponding opinion word in reviews. It is thus possible to simultaneously model the aspect terms and their related opinion words to detect latent aspects and opinions from reviews. Additionally, one limitation of existing topic models is that the correspondence between detected semantic opinion variables and real opinion orientations (e.g., positive or negative) is not explicit [TM08b, ME11]. To alleviate the problem, we exploit opinion prior information based on a sentiment lexicon in the modeling framework.

Therefore, we introduce a new supervised joint topic model called supervised joint aspect and opinion model (SJAOM) for the problem. SJAOM incorporates the overall ratings as supervision data, and simultaneously models pairwise aspect terms and opinion words in each review. One key advantage of SJAOM is that it can jointly identify the semantic aspects and opinions that are indicative of the overall ratings of customer reviews.

1.3 Summary of Contributions

We have made contributions to opinion mining field. We summarize our major contributions in this thesis as follows:

- Extracting aspect terms and opinion words via corpus statistics association based bootstrapping method (ABOOT).

  (i) We address a problem of extracting domain-specific aspect terms and opinion words that appear in reviews, which is essential and practical in fine-grained aspect-based opinion mining.

(ii) We develop a unified bootstrapping framework based on a small list of known seeds, which exploits word-to-word dependency relations to discriminate valid aspect terms and opinion words from invalid ones.

(iii) We propose to employ corpus statistics association models but not to rely on syntactic dependency rules to identify the dependency relations for aspect and opinion extraction. This is because syntactic dependency rule based methods tend to suffer from: 1) the limited coverage of the manually developed rules, and 2) the informal content of real-life reviews which typically contain grammatically incorrect sentences or misspellings.

(iv) The experimental results show the proposed ABOOT method, in particular LRTBOOT, consistently outperforms the benchmark methods. The consequence of our discovery is far reaching: starting with just one aspect seed, typically the domain word, e.g., "cellphone" for cellphone review domain, LRTBOOT can automatically extract a large number of domain-specific aspect terms and opinion words from the given review corpus without any supervision knowledge other than the single domain seed. This makes our proposed association-based bootstrapping approach powerful and effective for practical aspect term and opinion word extraction in opinion analysis of customer reviews.

- Identifying implicit aspects for the opinion words devoid of explicit aspects via cooccurrence association rule mining method (coARM).

  (i) Little work has been done on implicit aspect identification in opinion mining. We proposed a new two-phase cooccurrence association rule mining approach (coARM) for the problem.

(ii) We formulate the problem as a specialized type of association rule mining. An association rule is generated based on a cooccurrence matrix between opinion word group and aspect term group, where rule antecedent comes from the opinion word group, while rule consequent comes from the aspect term group.

(iii) To the best of our knowledge, this is the first work that conducts quantitative evaluations based on real-world reviews for implicit aspect identification. The experimental results demonstrate the benefits of coARM over baseline methods.

- Modeling and mining customer reviews via supervised joint aspect and opinion model (SJAOM).

  (i) We deal with the problem of identifying semantic aspects and opinions on the aspects, as well as predicting overall ratings of reviews in a unified modeling framework.

  (ii) Most of existing topic modeling approaches are unsupervised. We exploit the overall review ratings as supervision data and introduce a new supervised joint topic model SJAOM.

  (iii) To the best of our knowledge, our approach is the first to use supervised joint topic models for opinion mining problem. One key advantage of SJAOM is that it can discover the semantic aspects and opinions which are predictive of overall ratings of reviews.

  (iv) Given new unlabeled reviews, we form the prediction of overall ratings via a carefully-designed regression procedure on the detected aspects and opinions in the reviews. This is different from almost all existing topic modeling approaches, which typically formulate the overall opinion prediction as a classification problem, and then simply classify each review based on the per-document opinion distributions.

(v) Experimental results show the improved effectiveness of SJAOM over six existing major models on real-world reviews.

1.4 Thesis Organization

This thesis is organized as follows. In Chapter 2, we conduct a detailed literature survey on opinion mining and aspect-based opinion mining. We also provide the definitions of the terminologies commonly used in opinion mining field. Chapter 3 investigates the problem of aspect term and opinion word extraction. Chapter 4 studies the problem of identifying implicit aspects for the opinion words devoid of explicit aspect terms. In Chapter 5, we focus on modeling customer reviews with aims at identifying aspects and opinion orientations as well as predicting the overall ratings of the reviews in a unified framework. We introduce a novel SJAOM model for this problem. In Chapter 6, we conclude the thesis with a summary of our work and also provide several promising future directions.

Chapter 2

Literature Survey

In this chapter, we first survey opinion mining research at different levels of granularity. Then, we conduct a survey of existing work on fine-grained aspect-based opinion mining. Many different approaches have been developed to address aspect-based opinion mining problems. We group the existing approaches into four categories: 1) Supervised learning approaches, 2) Linguistic approaches, 3) Corpus statistics approaches, and 4) Topic modeling approaches. Following that, we discuss implicit aspect identification problem.

2.1 Opinion Mining

2.1.1 Terminology

There are several terminologies commonly used in opinion mining. We describe each of them in detail as follows.

Opinion Mining: Opinion mining, also known as sentiment analysis [YNBN03, PL08], opinion analysis [RPW06, SC08], etc., is a field of study that analyzes people's opinions, sentiments, judgements, evaluations, and attitudes towards entities like products or services [Liu12]. The opinions and sentiments expressed in textual reviews can be analyzed at different levels of granularity, such as document level [PLV02], sentence level [HW00], and phrase (word) level [HM97].

Opinion Orientation: An opinion orientation, or opinion for short, simply indicates the sentimental polarity, or the degree (strength) of satisfaction demonstrated in a review text. Positive orientation indicates a desirable state, e.g., good and nice, while negative orientation indicates an undesirable state, e.g., bad and poor. Opinion orientation is also known as semantic orientation [HM97, TL03], sentiment orientation or sentiment [BHV04], opinion polarity [WWH05], sentiment polarity [PL08], etc.

Opinion Word: An opinion word refers to the word used to express opinion and sentiment in a review. It can appear as an adjective, verb, adverb, or even noun in review sentences [Wie00, RWW03, TL03, CVXS06]. It is also called sentiment word [YNBN03], opinion expression [JHS09], etc. For example, an opinion word "beautiful" can be recognized from this cellphone review "The exterior is very beautiful".

Aspect: An aspect refers to a distinct ratable facet of an entity (product), and it can be a semantic attribute, component, function, or feature of the entity, e.g., "screen" of a cellphone.

Aspect Term: An aspect term refers to a particular term used to express an aspect. For example, the specific aspect terms "screen", "LCD", and "display" are all indicative of the unique semantic aspect "screen" on cellphone review domain. Aspect term consists of explicit aspect term and implicit aspect term.

Explicit Aspect Term Explicit aspect term, or aspect term for short, indicates the term that is explicitly mentioned in customer reviews. It typically appears as noun or noun phrase in the review sentences and is also known as explicit aspect expression or explicit aspect [Liu12], explicit feature [Liu10], feature [HL04a], feature term [YNBN03], or opinion feature [HL04b].

Implicit Aspect Term: Implicit aspect term, or implicit aspect, is opposite to the explicit aspect term defined above. It indicates the aspect term that does not appear but is implied by an opinion word in reviews. It is also known as implicit feature in opinion mining domain [Liu10]. For instance, "It is very beautiful!", the opinion word "beautiful" implies an implicit aspect (term) "appearance" that has not been mentioned in this cellphone review.

In the following, we will first describe existing work on opinion analysis at document (review), sentence, or phrase (word) level, respectively.

2.1.2 Document-level Opinion Analysis

Document-level opinion analysis has been well studied. Generally, the goal is to predict the overall opinion orientation expressed in a whole review document. It can be formulated as either opinion classification [PLV02, Tur02] or opinion rating prediction [PL05] problem. As an opinion classification problem, the task is to classify whether the review document expresses a positive or negative opinion orientation as a whole. As for the opinion rating prediction problem, the task is to predict the overall numerical rating of a review.

Opinion Classification

In text mining filed, supervised learning algorithms such as Naive Bayes, maximum entropy, and support vector machines have be shown effective for text categorization problems [Lew98, BPP96, Joa98]. Pang et al. [PLV02] evaluated the effectiveness of applying the three supervised learning methods for determining whether a review document is positive or negative. They found that standard machine learning techniques achieved good results in comparison to human-generated baselines. The support vector machines (SVM) based on unigram term weighting performed the best for sentiment classification task. However, machine learning methods did not perform as well on sentiment classification as on traditional topic-based text categorization.

When classifying the sentiment of a review as positive or negative, it is obvious that not all sentences in the review are equally discriminative or informative. To prevent a sentiment classifier from considering the irrelevant or even potentially misleading sentences, Pang and Lee [PL04] extended their previous work by integrating a sentence-level subjectivity detector with the document-level sentiment classifier. Specifically, they introduced a minimum cut based classification to identify the sentences in a review as either subjective or objective, and discarded the objective ones. They then used SVM and Naive Bayes models to classify the subjective portions of each review document, leading to the improved opinion analysis performance. Similarly, Yessenalina et al. [YYC10] proposed a two-level structural SVM for the document-level sentiment classification problem. The proposed method first identifies the subjective sentences in a review document and then classifies the sentiment of the document based on only the extracted subjective sentences.

Ng et al. [NDA06] conducted the subjectivity analysis not at sentence level but at document level. In particular, they first identified whether a given text document is a review or not by using a SVM classifier trained on unigram features. They then classified the sentiment orientation of each identified review document via another SVM classifier trained using linguistic knowledge sources in addition to unigram features.

As support vector machines algorithm has been shown effective for sentiment classification problem, much work has been done on developing different sets of data factors to further improve the opinion analysis performance. Besides the unigram data features (factors), Mullen and Collier [MC04] also studied a variety of diverse indicators including favorability measures and textual topic knowledge to learn a SVM model for document-level sentiment classification. Whitelaw et al. [WGA05] proposed to train a SVM sentiment classifier using the appraisal groups in addition to standard unigram word features. An appraisal group is represented as a set of attribute values in the task-independent semantic taxonomies based on appraisal theory. In order to predict the sentiments of movie reviews, Kennedy and Inkpen [KI06] trained a SVM classifier using a different list of data features, including positive and negative expressions, valence shifters (i.e., negations, intensifiers, and diminishers), as well as simple unigram features.

Based on a set of annotated training data, supervised sentiment classification methods may be tuned to work well on a domain, e.g., cellphone reviews. But the models often need retraining when applied to a different review domain. Moreover, annotating a high-quality set of review data for each domain is very expensive in practice. An alternative to the bottleneck is to develop unsupervised sentiment classification techniques, which have no need for the annotation review data to do sentiment analysis.

Typically, unsupervised learning approaches classify the overall opinion of a review document by aggregating the sentiment orientations of the individual parts such as opinion expressions mentioned in the document. The sentiment orientation of the opinion expressions can be identified via a corpus-based method or a dictionary-based method. In addition, the modifiers of opinion expressions, such as negations or intensifiers mentioned in the document, can be also considered for the opinion classification task.

Turney [Tur02] proposed a simple unsupervised statistical association analysis method to classify reviews as recommended (thumbs up) or not recommended (thumbs down). By employing a point-wise mutual information based method, The author first computed the semantic orientation of a given opinion phrase as the mutual information between the given phrase and the word "excellent" minus the mutual information between this phrase and the word "poor". Then, the author classified a review as recommended if the average semantic orientation of the opinion phrases contained in the review is positive, and vice versa. One limitation of this work lies in its reliance on an external search engine for computing the pairwise mutual information.

Taboada et al. [TBT+11] proposed a lexicon-based method to classify the sentiment of a text review. They relied on the opinion words defined in a sentiment lexicon as well as the modifying intensifications and negations to perform the sentiment classification task. Though domain-independent, their method may not work properly on large scale review corpus, as it is very likely that a great many opinion words of the corpus are not covered by the pre-compiled opinion lexicon. Similarly, Paltoglou and Thelwall [PT12] proposed an unsupervised lexicon-based method for sentiment analysis in social media like Twitter, MySpace, and Digg. They added an extensive list of linguistically driven functionalities to the method, including negation/capitalization detection, intensifier/diminisher detection, and emoticon/exclamation detection.

Dave et al. [DLP03] also proposed an unsupervised method for determining whether a review is positive or negative. They developed appropriate features as well as feature scoring methods based on information retrieval techniques. They then computed the overall sentimental score of a review document based on the scores of individual features covered by the document. Zagibalov and Carroll [ZC08] proposed an iterative approach to sentiment seed selection for unsupervised sentiment classification of each text review. They determined the overall sentiment of a review document based on the positive and negative zones in the document, where the zone means a sequence of words delimited by punctuation marks. The sentiment score of a zone is the sum of sentiments of all the sentiment seed words found in the zone.

Simply formulating the opinion classification as a supervised text categorization problem suffers from the painstaking requirement of collecting high-quality annotation review data in each domain. On the other hand, using background prior knowledge alone for opinion analysis cannot benefit from those state-of-the-art supervised learning algorithms. Therefore, semi-supervised learning methods could be a good try for the dilemma, which can take advantage of the background prior knowledge, for instance, in the form of sentiment words, abundant unlabeled data, as well as a few labeled review documents to learn a high-quality sentiment classification model with minimal human supervision.

Melville et al. [MGL09] introduced a machine learning approach which effectively combined background sentiment lexical knowledge with supervised learning. In particular, they fist constructed a generative model based on a lexicon of sentiment-laden words, and trained another model on the small set of labeled documents. The word distributions from the two models are then adaptively pooled to create a composite multinomial naive Bayes classifier that captures both sources of information. There are two advantages for the proposed approach: 1) By exploiting the prior lexical knowledge, they dramatically reduced the amount of training data required; and 2) By using some labeled documents, they can refine the background lexical knowledge, which in turn enables the trained model to effectively adapt to the review domain.

Li et al. [LHZL10] used two views, namely, personal and impersonal views, in a semi-supervised learning setting for sentiment classification of customer reviews. They first used an unsupervised bootstrapping method to automatically separate each review document into personal and impersonal views. They then exploited the two views in a supervised sentiment classification via an ensemble of individual classifiers generated based on each view. They finally introduced a cotraining algorithm to incorporate unlabeled data for semi-supervised sentiment classification problem.

Additionally, several other similar semi-supervised learning methods were developed for document-level sentiment classification problems. Beineke et al. [BHV04] improved review classification by using a semi-supervised scheme based on labeled documents, unlabeled documents, and human-provided information. They used an unsupervised semantic association method [Tur02] to effectively generate a new set of labeled review documents. They then trained a naive Bayes classifier based on the training data to classify review sentiments. Sindhwani and Melville [SM08] proposed a generalized framework for incorporating the lexical information as well as unlabeled data within standard regularized least square method for the sentiment prediction task. Tan et al. [TWC08] proposed a seme-supervised method that classifies the sentiment of each review without using labeled data. Li et al. [LZS09] proposed a non-negative matrix factorization approach for sentiment classification, which employs the lexical prior knowledge in the form of domain-independent sentiment laden terms, domain dependent unlabeled data, and a few labeled documents. He [He10] proposed a semi-supervised learning framework for sentiment classification task. He first constructed a classifier based on prior background knowledge to classify the sentiments of documents. Then He used the documents classified with high confidence as pseudo-labeled examples to automatically extract domain-specific data features. The recognized data features were finally used to train another classifier for sentiment prediction of reviews.

Opinion Rating Prediction

The aforementioned work on opinion analysis mainly focuses on the binary classification of reviews as positive or negative. Generally, it could be more meaningful and informative to infer a sentimental rating or score, for instance, in the form of 1 to 5 stars.

Pang and Lee [PL05] dealt with the sentimental rating inference problem, wherein rather than simply determining whether a review is thumbs up or thumbs down, instead, they identified reviewers' evaluations with respect to a multi-point scale (e.g., 1 to 5 stars). By formulating as a metric labeling [KT02] problem, and introduced a SVM based meta-algorithm for the rating analysis, which incorporates information about item similarity together with label similarity information.

To analyze numerical opinion ratings of reviews, Qu et al. [QIW10] first introduced a bag-of-opinions representation of document, and then developed a constrained ridge regression method to learn the score of each opinion phrase from the domain-independent corpora of rated reviews. They predicted the sentimental rating of a review by aggregating the scores of all opinion phrases in the review. The domain-dependent unigram information was also considered in this task.

In addition, Li et al. [LLJ+11] proposed a tensor factorization technique to incorporate the reviewer, product, and text features for review rating prediction. Goldberg and

Zhu [GZ06] presented a graph-based semi-supervised learning method to infer numerical review ratings. Their method takes advantage of the unlabeled reviews in the model learning process to improve the rating inference performance.

2.1.3 Sentence-level Opinion Analysis

Opinion analysis at sentence level focuses on analyzing the subjectivity or opinion orientation expressed in a review sentence. Sentence-level opinion analysis has different characteristics compared to the aforementioned document-level opinion analysis:

- In general, review sentences are short, and limited data features are available in each sentence. Standard supervised classification methods trained on such sentences may suffer from sparse data problem.

- The sentiments of individual sentences in a document are not independent of one another. It is common for people to express the same opinion (positive or negative) across sentences unless there is an indication of opinion change using expressions like "but" or "however".

- Subjective clues, such as opinion words and modifying words (intensifiers and negations), are predicative of the sentence subjectivity or sentiment.

Hatzivassiloglou and Wiebe [HW00] studied the effects of dynamic adjectives, semantically oriented adjectives, and gradable adjectives on a simple subjectivity classifier, and concluded that the adjectives are strong predictors of sentence subjectivity.

Yu and Hatzivassiloglou [YH03] recognized the subjective sentences from the factual ones by using different methods, namely, SimFinder-based similarity approach [HKH$^+$01], Naive Bayes classifier, and multiple Naive Bayes classifiers. They then proposed another unsupervised method to identify the sentiment of each recognised opinion sentence by aggregating the opinion orientations of individual opinion words mentioned in the sentence.

Wiebe and Riloff [WR05] proposed a semi-supervised learning framework to detect the subjectivity of a sentence. They first generated an initial set of training data using rule-base subjective and objective classifier [RW03]. Then, based on the initial training data, they used the AutoSlogTS algorithm [Ril96] to learn new discriminative subjective and objective extraction patterns. The recognized patterns were then used in the aforementioned rule-based classifier to generate new high-confidence training data. Finally, a naive Bayes based self-training method was trained on the data to separate the subjective sentences from the objective ones.

Zhao et al. [ZLW08] addressed the sentence sentiment classification by using a conditional random fields (CRF) based method, which captures the influence of contextual information among sentences on the task. Ganu et al. [GEM09] proposed a regression-based method for sentence rating prediction, which takes into account textual components of reviews, such as categorization topics and sentiments.

2.1.4 Phrase-level Opinion Analysis

Opinion analysis at phrase (word) level focuses on inferring the semantic orientations of opinion phrases and words. As described in the previous sections, many approaches to opinion analysis at document or sentence level begin with building a lexicon of opinion phrases or words with known semantic orientation. It is thus non-trivial to study effective techniques for phrase-level (word-level) opinion orientation inference. Existing approaches to the task can be roughly divided two categories: corpus-based methods and dictionary-based methods.

Hatzivassiloglou and McKeow [HM97] studied the problem of identifying the semantic orientations of adjectives using a corpus-baaed method. They first recognized semantic constraints from conjunctions (e.g., "and" or "but") on adjectives from a given corpus. They then used the constraints to construct a log-linear regression model to predict whether the two conjoined adjectives are of the same or different orientations. After that, they applied a clustering algorithm to separate the adjectives into two groups of different orientations. They finally exploited linguistic prior knowledge to classify the group of adjectives with higher average frequency as positive orientation. One shortcoming is that a good many adjectives that are mentioned in isolation in sentences are overlooked.

Turney and Littman [TL03] introduced an unsupervised method, which relies on corpus statistics association analysis, for inferring the semantic orientations of words. In particular, the semantic orientation of a given word is calculated from the strength of its association with a known set of positive paradigm words minus the strength of its association with a set of negative paradigm words.

Hu and Liu [HL04a] proposed a semantic dictionary based method to recognize valid opinion words as well as to infer the polarities of the words. They first provided a small list of seeds with known semantic orientations (30 common adjectives). They then iteratively expanded this list by discovering new synonymous and antonymous words of the given or extracted known adjectives on a semantic knowledge base called WordNet [MBF+90]. They classified the orientations of new recognized adjective opinion words based on the known orientations of the given seeds. One limitation of the method is that many adjectives appear several times in both positive and negative categories, which leads to conflicting semantic orientation inference results. Then, Kim and Hovy [KH04] extended their work by developing statistical methods to measure the polarity strength of each unknown adjective based on the synonymous words collected from the dictionary WordNet. They removed the sentiment-ambiguous adjectives but recognized only those with orientation strengths over a threshold. In addition, Esuli and Sebastiani [ES05] studied the word polarity inference problem via a learning approach based on the word gross representations extracted from a large dictionary.

Generally, the opinion orientations of words are context-dependent, meaning that the semantic orientations of prases/words may change across different context. Note that all of the aforementioned work does not consider the informative contexts of opinion phrases/words when inferring the semantic orientations.

Wilson et al. [WWH05, WWH09] presented a two-step learning method to recognize the contextual polarities of opinion expressions. In the first step, a neutral-polar classifier trained on different features was used to recognize the expression as neutral or polar in context. In the second step, another polarity classifier was developed to predict the contextual polarities of the recognized polar expressions.

Ding et al. [DLY08] proposed a holistic lexicon-based method to infer the context-dependent semantic orientations of opinion words. Their method takes advantage of different contextual information and evidences, such as negations, intra-sentence conjunctions, inter-sentence conjunctions, etc. Agarwal et al. [ABM09] first developed a list of prior polarity features, lexical features, and syntactic constituent features, etc. They then trained a logistic regression approach based on the features to do contextual phrase-level polarity analysis.

Yessenalina and Cardie [YC11] proposed a compositional matrix-space model based approach, which can capture the compositional effects required for accurate prediction of phrase-level sentiment. For instance, combining an adverb like "very" with a positive adjective like "good" produces an opinion phrase "very good", which expresses a stronger opinion polarity over the adjective alone.

Differently, Ku et al. [KHC09] proposed to rely on morphological and syntactic structures to infer the contextual polarities of words. They first recognized the morphological types of Chinese words by employing two models CRF and SVM. They then introduced a simple corpus statistics method to infer the opinion orientations of words in individual morphological classes. The syntactic dependency patterns of words in a sentence were then parsed and utilized to improve the word sentiment prediction in the context.

2.2 Aspect-based Opinion Mining

Opinion analysis at document (review), sentence, or phrase (word) level does not discover what exactly people like and dislike in customer reviews. In other words, it cannot identify the aspects commented on in reviews, and it thus fails to associate the predicted opinions with the specific aspects.

Clearly, identified opinions without the targeted aspects are of limited use in reality. Nowadays savvy consumers are no longer satisfied with just the overall opinion and rating about a product. They are eager to know why it receives the rating and which positive or negative aspects contribute to the final rating.

2.2.1 Problem Definition

Aspect-based Opinion Mining: Aspect-based opinion mining refers to mining and analyzing opinions and sentiments at the resolution of aspects as well as identifying fine-grained opinion orientations on the aspects. It is also known as aspect-based sentiment analysis [TNK10, Liu12].

Typically, there are two major tasks in aspect-based opinion mining:

- **Aspect Identification:** This task aims to discover what specific aspects of a reviewed entity (product) are evaluated in customer reviews.

- **Aspect-specific Opinion Prediction**: This task aims at identifying whether the opinions and sentiments towards detected aspects are positive or negative, or inferring the particular numerical opinion ratings on the aspects.

Both tasks of aspect identification and opinion prediction can be performed either in a unified framework [TM08b, JHS09, JO11], or in separate steps, for instance, identifying the particular aspects commented on in customer reviews followed by inferring the opinion orientations on the aspects [HL04a, PE05, BE10].

- **Input**

Apple iPhone 5
5 star: Everything great!! Attractive exterior, large display, long-lasting battery life.
4 star: I like it very much. It has a beautiful appearance and big screen, and its battery lasts longer.
4 star: The glass on the front is strong enough, and I never had to worry about getting scratched.
3 star: It has a bigger screen, very good-looking, but just too pricey…
3 star: Only newbies and wannabe Apple fans pay expensive price soon to be outdated!
1 star: It is really overpriced. DO NOT BUY IT!

Aspect-based Opinion Mining Algorithm

- **Output**

Aspect	Keywords (aspect terms/opinion words)		Rating
Exterior	exterior, appearance	attractive, beautiful, good-looking	5
Screen	display, glass, screen	large, strong, big, bigger	4
Battery life	battery life, battery	long-lasting, longer	4
Price	price	pricey, expensive, overpriced	1

Figure 2.1: How aspect-based opinion mining system works.

Figure 2.1 shows a simple example about how an aspect-based opinion mining system works for analyzing customer reviews. Typically, given a review corpus which contains six real-world reviews about the product "Apple iPhone 5" (collected from Amazon), the aspect-based opinion mining system takes it as input, and then automatically discovers and analyzes all the reviews in the corpus. The system lastly returns the opinion and sentiment summarization results as output, which contain a set of identified aspects commented on in reviews as well as a corresponding list of opinions/ratings expressed towards the aspects. For instance, an aspect "exterior" has been detected from the corpus, and the aspect groups a specific list of aspect terms "exterior" and "appearance". The identified opinion rating on the aspect is 5-stars/positive, which are expressed via a list of semantically related opinion words "attractive", "beautiful", and "good-looking".

Aspect-based opinion mining has been earlier known as feature-based opinion mining [HL04a, LHC05] or feature-based sentiment analysis [Liu10], where the terminology "feature" corresponds to the aspect.

In the following sections, we will describe existing techniques used for the development of such an aspect-based opinion mining system. We will first describe supervised learning approaches, followed by three unsupervised approaches, namely linguistic approaches, corpus statistics approaches, and topic modeling approaches. After that, we will discuss implicit aspect identification problem.

2.2.2 Supervised Learning Approaches

By modeling opinion mining tasks, such as aspect term and opinion word extraction and opinion orientation prediction, as sequential tagging problems, Jin et al. [JHS09] proposed a lexicalized Hidden Markov Models (HMM) based method. Different from hidden Markov models, their proposed lexicalized HMM naturally integrates linguistic features such as part-of-speech and surrounding contextual clues of words into the machine learning framework.

One limitation of this lexicalized HMM is that it is hard to integrate multiple data features in the framework. Li et al. [LHH+10] proposed a conditional random fields (CRF) [LMP01] based learning method for aspect-based opinion summarization of reviews. Similarly, by modeling the task as a sequence tagging problem, Jakob and Gurevych [JG10] employed a CRF-based method for aspect term extraction. In addition, authors also evaluated the CRF approach for aspect term extraction across different domains.

The performance of supervised tagging models highly rely on manually labeled training data. However, high-quality annotated training reviews are very tedious and expensive to come by in practice. Many different unsupervised approaches have been thus developed for aspect-based opinion mining problems.

2.2.3 Linguistic Approaches

Generally, semantic dependency relations naturally exists between aspect terms and opinion words. For example, in this cellphone review, “It has very beautiful appearance!”, the adjective opinion word “beautiful” is semantically related to the aspect term “appearance”. Mining such type of relations is very useful for identifying aspects and opinions from customer reviews.

Linguistic approaches rely on syntactic parsing to develop syntactic rules or dependency patterns for the aspect term and opinion word extraction as well as opinion polarity inference problems. The key idea is to employ syntactic rules or patterns to recognize the semantic dependency relations between each pair of aspect term and opinion word in every review sentence.

Qiu et al. [QLBC11] introduced a syntactic dependency rule based linguistic method to solve the opinion word and aspect term extraction from product reviews. First of all, they manually create eight syntactic rules based on the dependency grammar [Tes59]. They used the dependency rules to discover the pairwise semantic relations on the given review corpus. The identified dependency relations were then employed to iteratively locate the aspect terms and the corresponding opinion words from individual sentences. The accurate identification of relations is thus essential to the approach.

A semantic role labeling based linguistic approach was used to identify the aspect terms and the associated opinion words expressed in news media text in [KH06]. Qiu et al. [QWB+08] developed a linguistic method to extract aspect terms commented on in reviews as well as to predict the contextual opinion orientations on the aspect terms. Wu et al. [WZHW09] proposed a linguistic method based on phrase dependency parsing to recognize the aspect terms, opinion expressions, as well as the relations between them in reviews.

Generally, linguistic approaches are domain-independent, in the sense that the syntactic rules or dependency patterns developed in a domain can be applied across domains. However, the approaches tend to suffer from the problems: 1) the limited coverage of the manually-defined syntactic rules, and 2) the informal characteristics of real-life reviews, which typically contain colloquial content or grammatically incorrect sentences. One alternative to the methods is relying on corpus statistics analysis for fine-grained opinion mining tasks.

2.2.4 Corpus Statistics Approaches

Corpus statistics analysis approaches rely on mining frequent corpus statistics patterns to address the aspect term and opinion word extraction as well as the opinion orientation prediction problems. The approaches are somewhat resistance to the colloquial nature of customer reviews, given a suitably large review corpus.

Hu and Liu [HL04a] developed a corpus statistics method for mining and summarizing online customer reviews. They used an association rule mining [AIS93] approach to discover frequent nouns and noun phrases as candidate aspect terms. They then applied compactness and redundancy pruning to filter out the noisy aspects. One limitation of the method is that many aspect terms which are frequent but invalid are extracted incorrectly from product reviews. Popescu and Etzioni [PE05] proposed to employ a point-wise mutual information based corpus statistics method to evaluate each extracted aspect candidate, and pruned the frequent but invalid potential aspects, leading to improved performance. Once the aspect terms are identified, the adjacent adjectives of the aspects are located as the modifying opinion words. The semantic orientations of the opinion words can be then inferred by employing corpus statistics association methods [TL03].

Ku et al. [KLWC05, KC07] recognized topical words (i.e., aspect terms) from online reviews using a simple statistical frequency method. Then, they retrieved relevant review sentences using the detected topical words, and classified the sentence sentiments using pre-compiled opinion lexicons. A topic-based (aspect-based) opinion summary of reviews can be generated.

Zhang et al. [ZLLOS10] first extracted candidate aspect terms from reviews by employing syntactic patterns "part-whole" and "no" plus the syntactic rules defined in the work [QLBC11]. They then employed a hyperlink-induced topic search (HITS) algorithm to evaluate all the extracted aspect candidates, resulting in improved performance.

Xu et al. [XLL+13] proposed a two-stage method to extract opinion words and aspect terms from online reviews. In the first stage, they developed a graph walking algorithm to recognize the aspect and the related opinion candidates. In the second stage, they utilized a semi-supervised classifier to refine the extracted potential aspects and opinion words.

2.2.5 Topic Modeling Approaches

It is common that different people use different particular aspect terms to express the same semantic aspect. For instance, the terms "screen", "LCD", and "display" all indicate the single cellphone aspect "screen". The aforementioned supervised learning approaches, linguistic approaches, and corpus statistics approaches do not group the extracted synonymous aspect terms. There is thus redundancy in the opinion mining and summarization results. A separate step of categorization or clustering [GZG+09, ZLXJ10, ZLXJ11] may be used to generate much concise and meaningful results, unfortunately, which tend to additionally lead to the accumulation of errors.

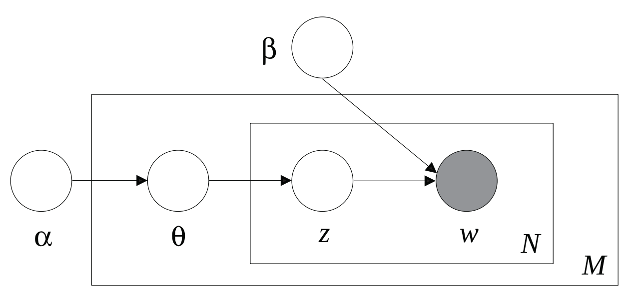

Figure 2.2: Graphical model representation of LDA [BNJ03].

Probabilistic topic models have been shown effective for discovering semantic topical structure of text data [Hof99, BNJ03]. Latent Dirichlet allocation (LDA) [BNJ03] is a basic generative topic model, which is based on the idea that documents are modeled as mixtures over a set of semantic topics, where each semantic topic is represented as a probability distribution over words. Figure 2.2 shows the graphical model representation of LDA. The plates indicate replicates. The outer plate means a collection of textual documents, while the inner plate represents the repeated choices of topics and words within a document. Different extensions on this basic LDA model has been developed to jointly detect the semantic aspects/topics and opinions on the aspects/topics from textual reviews.

Based on the basic LDA, Titov and McDonald [TM08a] introduced an unsupervised multi-grain LDA model (MG-LDA) for extracting ratable aspects from reviews. Then, they extended the MG-LDA model by adding an opinion rating prediction module. They introduced a multi-aspect sentiment model (MAS) [TM08b] to identify latent aspects/topics and the ratings on the aspects from customer reviews. The assumption of MAS that individual aspect-rating pairs are present in review data may lead to its limited use in reality. This is because a large quantity of online reviews are not annotated with aspects as well as aspect-specific ratings by customers.

Lin and He [LH09] extended the standard LDA model by adding a sentiment layer and proposed an unsupervised joint sentiment and topic model (JST) for opinion analysis. One advantage of JST is its ability to jointly detect the latent topics and sentiments from customer reviews. JST can be used to analyze reviews at various levels of granularity, including document-level opinion classification and aspect-based opinion analysis. Then, Lin et al. [LHER12] extended the JST model by incorporating sentiment priors and introduced a new weakly supervised joint sentiment-topic model. The supervised data comes from a domain independent pre-compiled sentiment lexicon. The weakly supervised nature of the model makes it easily portable to different domains for opinion classification [HLA11]. One limitation of the weakly supervised JST is that the detection for the hidden topics or aspects is unsupervised, negating the supervised advantage.

Wang et al. [WLZ10] proposed a generative regression model to address the latent aspect rating analysis problem (LARA). They first detected semantic aspects using a straightforward bootstrapping algorithm. They inferred the opinion ratings on the aspects using a probabilistic regression model. One limitation is that a set of aspect keywords must be pre-defined for the aspect detection. To solve this issue, they then developed a new latent aspect rating analysis model, which can perform the tasks of identifying semantic aspects and ratings in a unified framework, and does not need to pre-specify aspect keywords [WLZ11]. One challenge that remains in this new model is the aspect segmentation capability, which may provide inaccurate segments for the subsequent aspect-based rating analysis task.

In addition, Jo and Oh [JO11] proposed an aspect and sentiment unification model (ASUM) to detect sentiments towards different aspects in a unified framework. The assumption that each review sentence contains exactly one aspect/topic is often violated when modeling and mining real-life long and complicated reviews. Moghaddam and Ester [ME11] proposed an interdependent LDA model (ILDA) to jointly detect latent aspects (topics) and sentiments from textual reviews. One shortcoming of this unsupervised ILDA is that the correspondence between latent sentiment variables and real-world sentiment orientations (e.g., positive and negative) is not explicit.

2.2.6 Implicit Aspect Identification

When people share their comments on a product, they often mention opinion words but miss out the corresponding aspects of the products in reviews. For example, "It is really beautiful, buy one please.", a positive opinion has been expressed via the opinion word "beautiful" on the cellphone aspect "appearance", which is not mentioned explicitly in this review.

An implicit aspect refers to the aspect that does not appear but is implied by an opinion word in a review. The task of implicit aspect identification is to infer specific aspects for the opinion words that do not have corresponding explicit aspect terms in individual reviews.

Liu [Liu10] suggested to address the problem of implicit aspect inference by mapping opinion words, also known as aspect indicators, to their underlying aspects based on a manually-compiled semantic dictionary. "However, it is not clear whether this is an effective approach as little research has been done on the problem" [Liu10].

To deal with aspect-based opinion mining problem, Su et al. [SXG+08] developed a mutual reinforcement clustering method to construct a hidden sentiment association set between the aspect and opinion word groups. The method may be potentially useful for the implicit aspect identification, but no evaluations have been conducted for this problem.

2.3 Summary

In this chapter, we survey the related work to opinion mining as well as fine-grained aspect-level opinion mining.

Opinion mining, also known as sentiment analysis, has been studied for years. Generally, opinions and sentiments expressed in reviews can be analyzed at different levels of granularity. Opinion analysis at phrase/word level, which focuses on inferring the semantic orientations of individual opinion phrases/words, can be incorporated as a building block of opinion analysis at a higher level, such as sentence or document level. Supervised learning methods, unsupervised lexical knowledge based methods, as well as semi-supervised learning methods have been developed to predict the opinions and sentiments at different resolutions.

However, opinion mining at the document, sentence, or phrase/word level does not identify what specifically people like and dislike in reviews. It thus fails to associate the recognized opinions to the particular aspects commented on in reviews. In general, identified opinions and sentiments devoid of their corresponding aspects are of limited use in reality.

In this thesis, we focus on the problem of aspect-based opinion mining of customer reviews. We have made major contributions to the field: 1) We propose a generalized corpus statistics association based bootstrapping method (ABOOT) to extract domain-specific aspect terms and opinion words that appear in reviews; 2) We propose a two-phase cooccurrence association rule mining method (coARM) to identify implicit aspects for the opinion words devoid of explicit aspects; and 3) We developed a novel supervised joint aspect and opinion model (SJAOM) for modeling customer reviews with aims at identifying semantic aspects and opinions on the aspects as well as predicting overall ratings of the reviews. In subsequent chapters, we will describe the problems and our novel approaches proposed for them in detail.

Chapter 3

Extracting Aspect Terms and Opinion Words via Association-based Bootstrapping

3.1 Introduction

Opinion analysis at document, sentence, or phrase (word) level does not discover what people like and dislike in reviews. In other wards, it cannot extract the aspects commented on in customer reviews and thus fails to associate the identified opinions with the aspects in the reviews. In reality, the recognized opinions and sentiments without their corresponding opinionated targets and aspects are of limited value.

Aspect-based opinion mining has attracted extensive attention recently. In this chapter, we focus on two essential tasks in this fine-grained opinion analysis of customer reviews: 1) extracting specific aspect terms commented on in reviews; and 2) extracting opinion words associated with the aspects. In opinion mining, an aspect term refers to the term used to express the attribute, component, function, or feature of a reviewed entity (product), which typically appears as noun or noun phrase in reviews, while an opinion word refers to the word used to demonstrate the subjectivity, opinion, sentiment, and evaluation towards an aspect.

Generally, semantic dependency relations naturally exist between aspect terms and opinion words, even among aspect terms or opinion words themselves in reviews. For example, the opinion word "beautiful" is often used to modify the aspect "appearance" in cellphone reviews.

We observe that the semantic dependency relations between each pair of words can be quantified and the results can be exploited to discriminate the valid aspect terms and opinion words from those invalid ones. We propose to employ corpus statistics association model to discover the pairwise semantic dependency.

We introduce a generalized association based bootstrapping method (ABOOT) for identifying aspect terms and opinion words that appear in customer reviews. ABOOT starts with a small list of known seeds, and then iteratively recognizes more and more valid aspect terms and opinion words from the given review corpus by employing corpus statistics association analysis. Two particular statistical association models, likelihood ratio tests (LRT) [Dun93] and latent semantic analysis (LSA) [DDH90], are evaluated for computing the semantic dependency, leading to two instances of this ABOOT method, namely, LRTBOOT and LSABOOT.

We evaluate LRTBOOT and LSABOOT using real-life reviews from cellphone and hotel domains. Experimental results show the benefits of our methods over benchmark methods. In particular, one aspect seed is all that is needed for LRTBOOT to outperform other methods. This seed can be simply the domain word, e.g., "cellphone" for cellphone review domain. The consequence of our discovery is far reaching: starting with just one aspect seed, typically the domain word, LRTBOOT can automatically extract two large sets of domain-specific aspect terms and opinion words from the given review corpus without supervision data.

We summarize our contributions in this work as following:

- We address an essential problem of extracting aspect terms and opinion words from customer reviews.

- We develop a generalized unified bootstrapping framework which exploits word-to-word dependency relations to solve the problem.

- We propose to employ corpus statistics association models to identify the semantic dependency relations.

- The proposed association-based bootstrapping method achieves surprisingly good aspect and opinion extraction performance with only one aspect seed.

3.2 Aspect Term and Opinion Word Extraction

3.2.1 Overview

When people express their comments on an aspect, a certain group of opinion words will be frequently mentioned. For example, to show opinions on the aspect "appearance", they usually choose the certain list of opinion words like "beautiful", "pretty", and "lovely", etc. Semantic dependency relations thus naturally exist between opinion words and aspect terms in customer reviews. In addition, there are also dependency relations among aspect terms or opinion words themselves, since people often express their opinions on several related product aspects in a single review.

We consider three types of word-to-word semantic dependency/association relations for aspect term and opinion word extraction, in particular, *aspect to opinion* (AO or OA), *aspect to aspect* (AA), and *opinion to opinion* (OO). Note that if we use only the AO dependency in the extraction process, we may miss out many valid aspect terms as well as opinion words. For example, given a known aspect "appearance" in cellphone reviews, we can recognize a set of semantically related opinion words like "beautiful" and "good-looking" using the AO dependency. However, if we do not consider the AA dependency, we may miss out the valid aspects like "exterior" or "screen" in reviews. Likewise, given the extracted known opinion word "beautiful", we could fail to identify the frequently cooccurring opinion words like "pretty" and "large", if we do not use OO dependency. Thus, in addition to the AO dependency, We also utilize both AA and OO association relations, which are often overlooked in previous opinion mining research.

Syntactic dependency rule based linguistic methods may be useful for identifying the semantic dependency to extract aspect terms and opinion words from reviews. However, they tend to suffer from: 1) the limited coverage of the manually defined syntactic rules, and 2) the informal language of real-world raw customer reviews, which typically contain misspellings and grammatically incorrect sentences.

We observe that the semantic dependency patterns between individual pairs of words can be quantified, and the results can be employed to discriminate valid aspect terms and opinion words from those invalid ones. We thus propose to employ corpus statistics association model to measure the pairwise semantic dependency relations.

According to our experiments, candidate aspect terms (opinion words) that are strongly associated with valid known aspects tend to be true aspects (opinions), while candidate aspects (opinions) that are strongly associated with invalid aspects tend to be false aspects (opinions). In other words, without ground truth in the form of known seeds, corpus statistics association approaches may lead to many frivolous aspect terms or opinion words. Thus, we propose to start out with a manually crafted list of domain-specific aspect seeds (A term is domain-specific, meaning that it is specific to the domain). We then iteratively enlarge this initial ground truth seed set by adding newly identified valid aspect terms and opinion words that are strongly associated with the known items in the set.

The validated seed set is like an exclusive invitation-only finals club, where a new member (candidate aspect term or opinion opinion) is inducted into the club if and only if he or she is well-known (strongly associated) to the existing members of the club (identified aspect terms or opinion words in the set). The initial club founders (seeds) thus play an essential role, and they need to "know" the relevant future members (aspect terms and opinion words), who would bring in yet more and more members to the club.

Therefore, we propose a generalized corpus statistics association based bootstrapping method (ABOOT) for aspect term and opinion word extraction from customer reviews. ABOOT first extracts a set of candidate aspect terms and a set of candidate opinion words from a given review corpus. Then, based on the initial set of aspect seeds, it identifies two sets of valid aspect terms and opinion words that have strong AA and AO associations with the aspect seeds, respectively. Next, based on the extracted known opinion word set, ABOOT again identifies new valid opinion words and aspect terms via OO and AO associations, respectively. The extraction process is performed iteratively until there are no new aspect terms or opinion words identified. Upon termination, we can extract a large number of valid aspect terms and the associated opinion words.

In the next section, we will describe the association-based bootstrapping algorithm in detail.

3.2.2 Association-based Bootstrapping Algorithm

Given a review corpus $\mathcal{C}$, we first need to generate two candidate sets of aspect terms and opinion words, from which valid aspect terms and opinion words are identified via mining corpus statistics association.

Aspect term typically occurs as noun (noun phrase) with the *subject* or *object* pattern in a review sentence. Simply selecting nouns or nouns phrases as aspect candidates gives good coverage (recall), but comes at the expense of letting in too many noisy candidates, which may negatively influence the subsequent aspect term extraction process.

We thus locate the nouns (noun phrases) with *subject* or *object* syntactic patterns in individual sentences of the corpus $\mathcal{C}$ to generate a candidate aspect term set $\mathcal{CA} = \{ca_1, \ldots, ca_i, \ldots, ca_M\}$ (M: set size). Next, we simply extract all adjectives and verbs in the review corpus $\mathcal{C}$ to form a candidate opinion word set $\mathcal{CO} = \{co_1, \ldots, co_j, \ldots, co_N\}$ (N: set size).

Table 3.1: Association matrices for bootstrapping aspect and opinion extraction.

Matrix	Entry	Association Type
M_{AO}	$AM(ca_i, co_j)$	*aspect to opinion* (AO)
M_{AA}	$AM(ca_{i_1}, ca_{i_2})$	*aspect to aspect* (AA)
M_{OO}	$AM(co_{j_1}, co_{j_2})$	*opinion to opinion* (OO)

Based on the two sets of candidate aspect terms and opinion words, we then compute a pairwise association matrix for each of the aforementioned dependency relations. As shown in Table 3.1, M_{AO}, M_{AA}, and M_{OO} represent AO (OA), AA, and OO association matrices, respectively. Let AM denote a generalized corpus statistics association model used to compute the pairwise association. Accordingly, the $AM(ca_i, co_j)$ ($ca_i \in \mathcal{CA}$, $co_j \in \mathcal{CO}$), $AM(ca_{i_1}, ca_{i_2})$ (ca_{i_1}, $ca_{i_2} \in \mathcal{CA}$), and $AM(co_{j_1}, co_{j_2})$ (co_{j_1}, $co_{j_2} \in \mathcal{CO}$) are AO, AA, and OO association scores estimated using the association model AM on the given review data $\mathcal{C}$.

Taking the association matrix M_{AO} for instance, to calculate the pairwise word association score $AM(ca_i, co_j)$ of the matrix, we first need to obtain the corpus occurrence statistics related to candidate aspect ca_i and candidate opinion co_j. We then estimate the association score between the candidates ca_i and co_j based on their corpus statistics by applying the association model AM. We compute two other association matrices M_{AA} and M_{OO} in a similar manner.

Our generalized corpus statistics association based bootstrapping approach, ABOOT in short, is summarized in Algorithm 1. Some variables in the algorithm are defined as follows:

- $\mathcal{S}$: a manually annotated aspect seed set that is used to supervise the process of bootstrapping aspect and opinion extraction.

Algorithm 1 Corpus statistics association based bootstrapping algorithm.

Require: Review corpus $\mathcal{C}$, aspect seed set $\mathcal{S}$
Ensure: Identified aspect terms and opinion words
$\mathcal{CA} \leftarrow$ Extract a candidate aspect set from corpus $\mathcal{C}$;
$\mathcal{CO} \leftarrow$ Extract a candidate opinion set from corpus $\mathcal{C}$;
$\mathcal{A} \leftarrow \mathcal{S}$;
$\mathcal{O} \leftarrow \varnothing$;
repeat
  for *each known aspect term* t *in* $\mathcal{A}$ **do**
    for *each candidate aspect ca in* $\mathcal{CA}$ **do**
      if *(*$AM(a, ca) \geq aath$*) AND (*$ca \notin \mathcal{A}$*)* **then**
        Extract candidate *ca* as an aspect term;
        Remove candidate *ca* from set $\mathcal{CA}$;
      end if
    end for
    for *each candidate opinion word co in* $\mathcal{CO}$ **do**
      if *(*$AM(a, co) \geq aoth$*) AND (*$co \notin \mathcal{O}$*)* **then**
        Extract candidate *co* as an opinion word;
        Remove candidate *co* from set $\mathcal{CO}$;
      end if
    end for
  end for
  for *each known opinion word* o *in* $\mathcal{O}$ **do**
    for *each candidate opinion word co in* $\mathcal{CO}$ **do**
      if *(*$AM(o, co) \geq ooth$*) AND (*$co \notin \mathcal{O}$*)* **then**
        Extract candidate *co* as an opinion word;
        Remove candidate *co* from set $\mathcal{CO}$;
      end if
    end for
    for *each candidate aspect ca in* $\mathcal{CA}$ **do**
      if *(*$AM(o, ca) \geq aoth$*) AND (*$ca \notin \mathcal{A}$*)* **then**
        Extract candidate *ca* as an aspect term;
        Remove candidate *ca* from set $\mathcal{CA}$;
      end if
    end for
  end for
  Update $\mathcal{A}$ and $\mathcal{O}$ with identified aspects and opinions;
until *No new aspect terms or opinion words are identified*
return Identified aspect set $\mathcal{A}$ and opinion word set O

- $\mathcal{A}$: an aspect set that keeps track of the extracted aspect terms, initially $\mathcal{A} = \mathcal{S}$.

- $\mathcal{O}$: opinion word set that tracks the extracted opinion words.

- $AM(w_1, w_2)$: an association score estimated via an association model AM between word w_1 and word w_2.

- *aoth*, *aath*, and *ooth*: association thresholds for the AO (or OA), AA, and OO association relations, respectively.

In Algorithm 1, we first extract two candidate sets of aspects and opinions in line 1 and line 2. In line 3, the aspect set $\mathcal{A}$ is initialized with the annotated set of aspect seeds. From line 6 to line 19, we identify new aspect terms and opinion words, which have strong associations with the known extracted aspects in the set $\mathcal{A}$. We then extract more opinion words and aspect terms based on the extracted known opinion set $\mathcal{O}$ from line 20 to line 33. In line 34 of the algorithm, we update both known aspect and opinion sets $\mathcal{A}$ and $\mathcal{O}$ with the extracted aspect terms and opinion words. The bootstrapping process is performed repeatedly, and will be terminated until no new aspect terms or opinion words are identified. Finally, we identify valid domain-specific aspect terms and opinion words from the given review corpus $\mathcal{C}$.

We illustrate Algorithm 1 using a straightforward example in Figure 3.1. Given a sample cellphone review corpus $\mathcal{C}$ (which contains 4 reviews only), we first extract a candidate aspect set $\mathcal{CA}$ and a candidate opinion set $\mathcal{CO}$ from the corpus $\mathcal{C}$. We then compute pairwise word association scores by employing an association model AM, as shown at the lower portion of the figure. Given an aspect seed "screen" and a threshold thd =2.0, we can extract an aspect set $\mathcal{A}$ and an opinion set $\mathcal{O}$ by applying the ABOOT algorithm, as shown at the upper right portion of the figure.

1. The screen is really big, but the price is too expensive!	*CA*={screen, price, student}
2. The price is expensive, students don't buy it usually.	*CO*={big, expensive, buy, beautiful}
3. The screen is beautiful, however the price is not!	*S*={screen}
4. The screen is big and beautiful!	*thd* = 2.0
	A={screen, price}
	O={big, beautiful, expensive}

AM	screen	price	student	big	expensive	buy	beautiful
screen		2.5	0.5	3.0	1.5	0.5	3.0
price	2.5		1.0	1.5	3.0	1.5	1.5
student	0.5	1.0		0.5	0.5	1.0	0.5
big	3.0	1.5	0.5		2.0	0.5	2.0
expensive	1.5	3.0	0.5	2.0		1.0	2.0
buy	0.5	1.5	1.0	0.5	1.0		0.5
beautiful	3.0	1.5	0.5	2.0	2.0	0.5	

Figure 3.1: A working example of the ABOOT algorithm.

3.2.3 Association Models for Bootstrapping

As shown in Algorithm 1, the proposed generalized framework for bootstrapping aspect term and opinion word extraction is called ABOOT (Association-based Bootstrapping). Different pairwise word association models can lead to different instance approaches.

Generally, there are two groups of thoughts on estimating pairwise dependency relations: one is the tests for statistical significance, the other is the association measure. On the tests for statistical significance front, we choose the likelihood ratio tests to compute the semantic dependency between each pair of words, leading to likelihood ratio tests based bootstrapping, LRTBOOT for short. As for the association measure, we use latent semantic analysis plus cosine similarity, and thus lead to latent semantic analysis based bootstrapping, namely, LSABOOT. We note that other types of association models can be also incorporated in the generalized bootstrapping framework.

Likelihood Ratio Tests for Bootstrapping

We first describe the *likelihood ratio tests* (LRT) [Dun93] association model. LRT builds a contingency table of two words w_i and w_j, derived from corpus statistics, as given in Table 3.2, where $k_1(w_i, w_j)$ is the number of documents (reviews) containing both terms w_i and w_j; $k_2(w_i, \bar{w}_j)$ is the number of documents containing term w_i but not w_j; $k_3(\bar{w}_i, w_j)$ is the number of documents containing term w_j but not w_i; $k_4(\bar{w}_i, \bar{w}_j)$ is the number of documents containing neither w_i nor w_j.

Table 3.2: Contingency table derived from corpus statistics.

Corpus statistics	w_j	$\bar{w}_j$
w_i	$k_1(w_i, w_j)$	$k_2(w_i, \bar{w}_j)$
$\bar{w}_i$	$k_3(\bar{w}_i, w_j)$	$k_4(\bar{w}_i, \bar{w}_j)$

Based on the corpus statistics shown in Table 3.2, LRT computes the statistical association between words w_i and w_j by employing the following function:

$$-2log\lambda = 2[logL(p_1, k_1, n_1) + logL(p_2, k_2, n_2) - logL(p, k_1, n_1) - logL(p, k_2, n_2)], \quad (3.1)$$

where,

$L(p, k, n)$=$p^k(1-p)^{n-k}$;

n_1=$k_1 + k_3$; n_2=$k_2 + k_4$;

$p_1 = k_1/n_1$; p_2= k_2/n_2;

$p = (k_1 + k_2)/(n_1 + n_2)$;

The higher the quantity $-2log\lambda$, the greater the statistical association between words w_i and term w_j. We abbreviate this LRT based bootstrapping as LRTBOOT.

Latent Semantic Analysis for Bootstrapping

Given a word-by-document matrix that represents a corpus of documents, *latent semantic analysis* (LSA) [DDH90] applies *singular value decomposition* (SVD) to the matrix to analyze the latent semantic structure of the text corpus.

In particular, by applying SVD we can decompose the word-by-document matrix into a product of three matrices:

$$X = LVR', \tag{3.2}$$

where L and R are the left and right singular matrices, and V is a diagonal matrix of *singular values*.

Let r be the rank of the raw matrix X. We select a value $k \ll r$. Let V_k denote the diagonal matrix generated by choosing the top k singular values from the matrix V, and let L_k and R_k be matrices generated by selecting the corresponding columns from the matrices L and R, respectively. We thus obtain a reduced matrix X_k by multiplying the three new matrices:

$$X_k = L_k V_k R_k'. \tag{3.3}$$

The matrix X_k is the best low rank (k) approximation to the original matrix X, which minimizes the *Frobenius norm* [GVL96] in the form:

$$\| E \|_F = \sqrt{\sum_{w=1}^{W} \sum_{d=1}^{D} |e_{wd}|^2}, \tag{3.4}$$

where $E = X$ - X_k, e_{wd}: element of matrix E, W: word dictionary size, and D: corpus size.

In the new latent space, we measure pairwise word associations via cosine similarity of the corresponding row vectors of the "smoothed" matrix X_k. We abbreviate this LSA based bootstrapping approach as LSABOOT.

3.3 Experiments

In this section, we describe and discuss the experimental results of aspect term and opinion word extraction on real-world customer reviews.

3.3.1 Experimental Setup

We evaluated the proposed LRTBOOT and LSABOOT methods against several benchmark methods. The BASELINE approach, which just uses the given aspect seeds as extraction results, was used only for aspect extraction evaluation. DP (*double propagation*) extracts iteratively aspect terms and opinion words by using the syntactic dependency relations identified via eight manually-defined dependency parsing rules [QLBC11]. DPHITS uses *hyperlink-induced topic search* algorithm (HITS) to validate potential aspect terms and opinion words recognized by DP as well as two additional dependency patterns of *part-whole* and *no* [ZLLOS10].

We first evaluated the aspect term extraction. In particular, we plotted the performance curve of precision versus recall, followed by the best aspect term extraction performance in F-measure. We also plotted the F-measure against the number of aspect seeds for each review domain. Next, we evaluated the opinion word extraction of the proposed methods LRTBOOT and LSABOOT against benchmark methods.

In our experiments, we used aspect terms but not opinion words as initial ground truth seeds, this is because opinion words are not as domain-specific as aspect terms. A word is domain-specific in the sense that it is specific to the domain.

3.3.2 Data Sets

We evaluated our proposed methods using real-life reviews from cellphone and hotel domains. Note that the proposed methods is language independent and can be easily extended to different language based reviews for aspect term and opinion word extraction.

The cellphone corpus comprises 7800 real-world reviews collected from a major Chinese website [1]. The hotel corpus contains 4900 reviews crawled from a famous Chinese travel portal [2]. All review documents in the two corpora were parsed using a Chinese language analysis tool named *Language Technology Platform* (LTP) [CLL10]. Some data statistics of the two corpora are given in Table 3.3.

Table 3.3: Some statistics of review data sets.

Domain	Cellphone	Hotel
# of reviews in corpus	7800	4900
# of sentences in corpus	12,546	18,239
# of words in corpus	3,160,997	9,462,144
Average # of words/sentence	252	519
Average # of words/review	405	1931

For the generation of a golden standard, we randomly selected 508 reviews from the cellphone review corpus, and two undergraduate students independently labeled the aspect terms and opinion words that actually appear in the reviews. Note that the annotated aspect terms (or opinion words) are confirmed valid if they are marked by both annotators. If only one annotator mark it, a third person make the final judgement. We obtained an annotated set of 995 aspect terms as well as a set of 1258 labeled opinion words for the cellphone reviews. Similarly, we annotated 1013 aspect terms and 1206 opinion words from 206 randomly selected hotel reviews.

3.3.3 Aspect Term Extraction

We first evaluated the aspect term extraction performance of the proposed methods against the competitors on cellphone and hotel reviews.

[1] product.mobile.163.com
[2] www.lvping.com/hotels

Precision versus Recall

In Figure 3.2 and 3.3, we plotted the precision-recall curves of LRTBOOT, LSABOOT, and DPHITS as well as showed a single precision-recall point for each of DP and BASELINE on the cellphone and hotel review domains, respectively. To maintain fairness, the reported results for all the evaluated methods operate on the same set of ten aspect seeds for each domain. The choice of ten seeds seems reasonable because people should be able to come up with ten common aspect seeds for a given review domain.

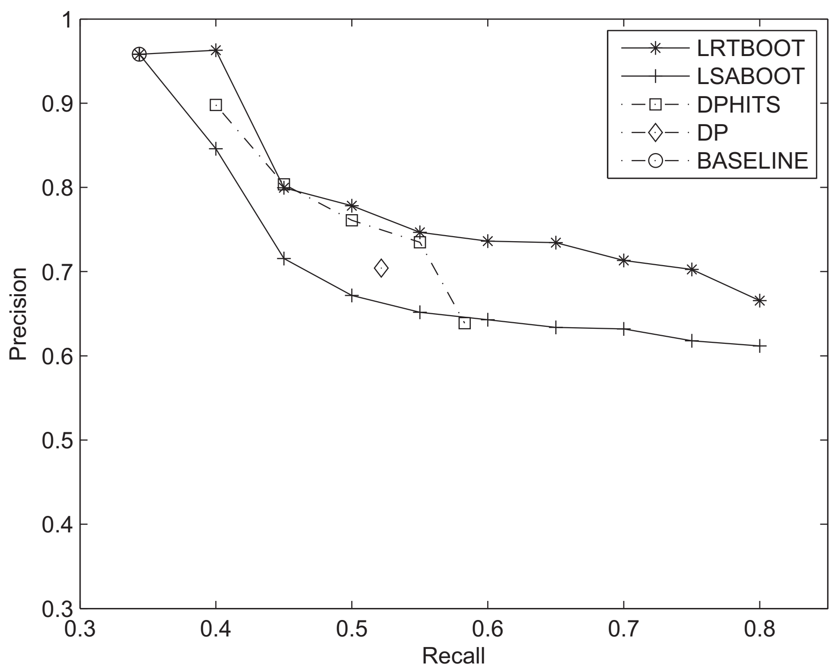

Figure 3.2: Cellphone aspect term extraction performance.

In Figure 3.2, LRTBOOT curve lies well above those of all competing methods on cellphone reviews. Despite starting off at the same precision at 35% recall, LSABOOT does not performs as well as LRTBOOT at all subsequent recall levels. The BASELINE method achieves a high precision of 95.8% at its single recall level of 35%. DP attains the precision of 70.4% at the 50% recall, which is 3.2% better than that of LSABOOT but 7.4% worse than LRTBOOT. The average precision of DPHITS is 76.7% across the recall levels from 40% to 60%, which is 6.1% better than LSABOOT, but 3.8% worse than that of LRTBOOT. The precision of DPHITS drops significantly to 63.87% as its recall approaches 60%, which is in the performance neighborhood of LSABOOT but 9.7% lower than LRTBOOT.

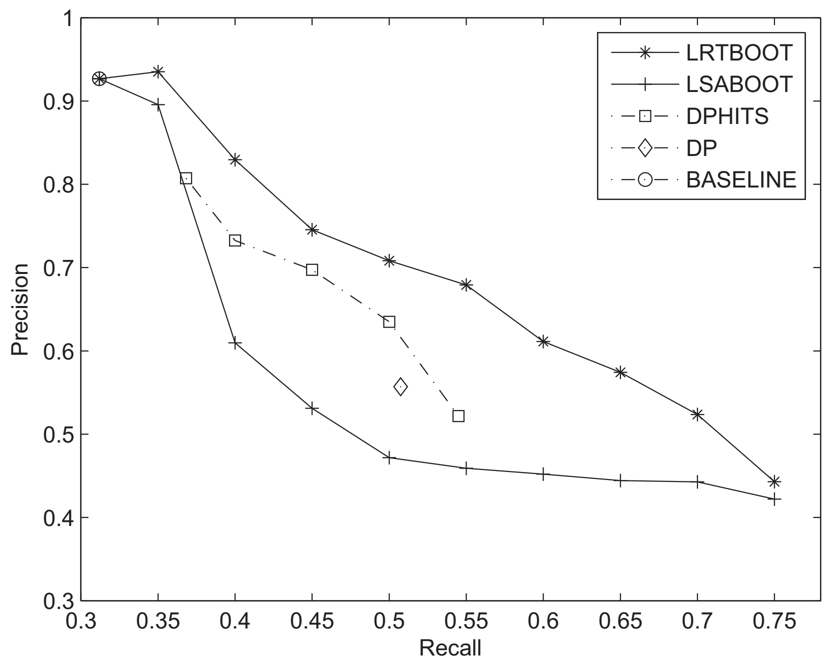

Figure 3.3: Hotel aspect term extraction performance.

Figure 3.3 plots the precision-recall curves on hotel reviews. LRTBOOT curve again lies far above the competitors for all recall levels, while LSABOOT performs worse compared with LRTBOOT. The BASELINE method achieves 92.7% precision at around 30% recall, which is the same as that of LRTBOOT and LSABOOT. DP achieves 55.7% precision at the single precision-recall point of approximately 50% recall, which is 8.5% better than LSABOOT, but 15.1% lower than LRTBOOT. Spanning the recall levels from 35% to 55%, the average precision of DPHITS is around 67.9%, which is 8.5% better than LSABOOT, but 10.1% worse than LRTBOOT.

Despite achieving good precision, the BASELINE method has extremely low recall values of 35% and 30% on cellphone and hotel review domains, respectively. DP enjoys relatively better coverage compared to the BASELINE, but performs worse than LRTBOOT and LSABOOT, leading to only 50% recall for both domains. DPHITS is marginally better than DP, but its coverage is still limited to at most 58% and 55% recall on the cellhone and hotel domains. Our proposed methods LSABOOT and LRTBOOT achieve much better coverage (higher recall) compared to all the competing methods which could not even cross the 60% mark. In practice, good performance (in precision) at a decent recall is more desirable.

Best Performance

Table 3.4: Best aspect term extraction performance in F-measure on cellphone reviews.

Method	Precision	Recall	F-measure
LRTBOOT	68.00%	79.40%	73.25%
LSABOOT	63.25%	81.11%	71.07%
DPHITS	71.00%	57.08%	63.29%
DP	70.42%	52.16%	59.93%
BASELINE	95.80%	34.37%	50.59%

Table 3.4 shows the best aspect term extraction results (in F-measure) on cellphone reviews. LRTBOOT is the overall winner in terms of robustness and absolute F-measure score. It achieves the largest F-measure of 73.25%, which is 9.96%, 13.32%, and 22.66% better than that of DPHITS, DP, and BASELINE, respectively. Thanks to the highest recall of 81.11%, LSABOOT is able to achieve the overall second best F-measure of 71.07%, which is 7.78%, 11.14%, and 20.48% better than DPHITS, DP, and BASELINE, respectively.

Though achieving the highest precision of 95.8%, the BASELINE method gives the lowest F-measure of 50.59%. This is expected since it only uses the ten seeds as the identified aspects, leading to the lowest recall of 34.37%. DP performs poorly compared to LRTBOOT and LSABOOT, and only results in the F-measure of 59.93%. DP locates the aspect terms by using eight manually defined syntactic dependency rules. It thus tends to suffer from: 1) the limited coverage of the rules, and 2) the informal language of customer reviews which typically contain many colloquial expressions or grammatically incorrect sentences. By using the HITS algorithm to evaluate the potential aspects extracted by DP and two new patterns, DPHITS gives better F-measure compared to DP, but still performs worse than LRTBOOT and LSABOOT.

Table 3.5: Best aspect term extraction performance in F-measure on hotel reviews.

Method	Precision	Recall	F-measure
LRTBOOT	57.43%	67.13%	61.90%
LSABOOT	45.34%	69.60%	54.91%
DPHITS	62.03%	52.42%	56.82%
DP	55.69%	50.74%	53.10%
BASELINE	92.67%	31.19%	46.68%

Table 3.5 shows the best aspect term extraction results on hotel reviews. LRTBOOT again achieves the overall best F-measure of 61.9%, which is 5.08%, 8.8%, and 15.22% better than that of DPHITS, DP, and BASELINE, respectively. LSABOOT achieves the F-measure of 54.91%, which is 1.81% and 8.23% better than DP and BASELINE, but 1.91% worse than DPHITS. Despite the best precision of 92.67%, BASELINE leads to the worst recall of 31.19% and the lowest F-measure of 46.68%. The experimental results again show that LRTBOOT improves the aspect term extraction performance compared to the benchmark methods DPHITS, DP, and BASELINE on the hotel review domain.

F-measure versus Seed Size

We studied the effect of seed set size on aspect term extraction for LRTBOOT and LSABOOT, as well as for benchmark methods DPHITS, DP, and BASELINE. To collect seeds, we simply ranked all potential aspect terms by descending corpus frequency for each review domain, and then manually selected up to fifty product aspects as seeds from the sorted list.

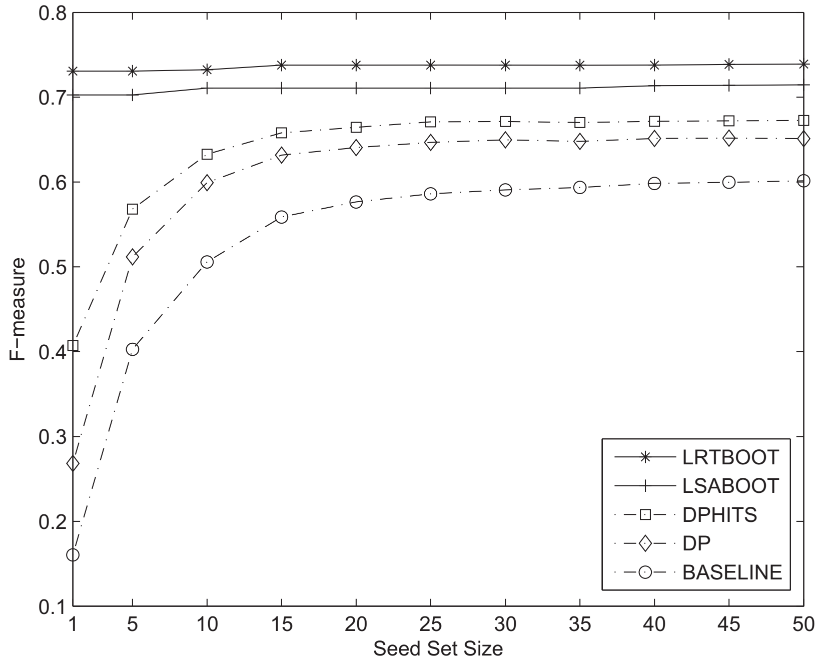

Figure 3.4: F-measure versus seed size for cellphone reviews.

Figure 3.4 plots the performance (in F-measure) versus top K seeds (up to fifty seeds) on cellphone reviews. Clearly, both LRTBOOT and LSABOOT outperform the competitors for all seed numbers from one to fifty. In particular, the average F-measure of LRTBOOT across all observations is 73.62%, which is 10.35%, 14.07%, and 21.13% better than that of DPHITS, DP, and BASELINE, respectively. The average F-measure of LSABOOT is 71.02%, which is 7.75%, 11.47%, and 18.53% better than that of DPHITS, DP, and BASELINE.

Furthermore, when increasing seeds from one to fifty, the F-measure scores of LRT-BOOT and LSABOOT remain almost constant. In other words, it does not depend critically on the seed set size. Remarkably, LRTBOOT achieves excellent performance with only one seed word, which is just the domain word "cellphone". A similar situation is also observed for LSABOOT. This is because the proposed methods benefit from corpus statistics association analysis for bootstrapping aspect term extraction, and given suitable association thresholds, candidate aspects that have strong associations with the domain word (top one seed) or the extracted known aspect terms or opinion words, will be identified sooner or later in several bootstrapping iterations. This is the major difference of our proposed method from the benchmark methods.

With growing the seed number, the F-measure scores of DPHITS and DP increase largely and begin to level off at twenty-five seeds. Clearly, seed set size results in big effect on the competing methods for aspect extraction. This agrees well with the observation that syntactic dependency rule based methods like DP and DPHITS tend to suffer from the limited coverage of the manually defined rules. Therefore, in order to achieve a decent performance, a relatively large seed set is needed for DPHITS and DP in practice. In contrast, just one word, i.e., the domain word, suffices for our proposed association-based bootstrapping method like LRTBOOT for domain-specific aspect term extraction.

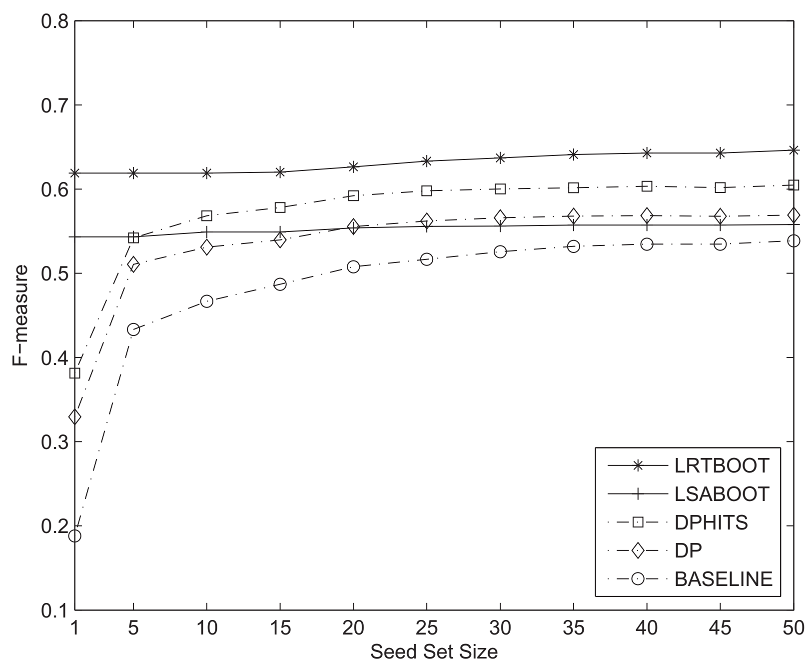

Figure 3.5: F-measure versus seed size for hotel reviews.

Likewise, we evaluated the aspect extraction performance versus seed set size on hotel reviews, as shown in Figure 3.5. Again the LRTBOOT curve lies well above that of DPHITS, DP, and BASELINE for all seed numbers, and this time LRTBOOT enjoys a slight visible improvement from fifteen to fifty seeds. The average F-measure of LRT-BOOT across all observations is 63.15%, which is 6.14%, 9.81%, and 15.28% better than that of DPHITS, DP, and BASELINE, respectively.

LSABOOT shows better performance for all seed sizes compared to the BASELINE method. It outperforms DP and DPHITS for seed sizes smaller than twenty and five, respectively. The average F-measure of LSABOOT across all seed numbers is 55.27%, which is 1.93% and 7.4% better than DP and BASELINE, but 1.74% lower than that of DPHITS. Different from cellphone reviews, hotel reviews are much longer, complicated, and contain a good many irrelevant personal anecdotes, which can make the exploration of hidden semantic structure of review data more challenging for latent semantic analysis model (LSA). This may explain the relatively lower F-measure performance of LSABOOT compared to DPHITS.

Moreover, LRTBOOT and LSABOOT show a high-level of consistencies across all seed sizes on the hotel reviews. The hotel results again validate our observation on the cellphone domain. We hence conclude that seed set size does not have significant influence on our proposed association-based bootstrapping approach. In practice, we can select only a domain word to apply LRTBOOT to the real-life reviews in the domain for aspect extraction with minimal supervision.

3.3.4 Opinion Word Extraction

We evaluated opinion word extraction for LRTBOOT, LSABOOT, and benchmark methods DPHITS and DP on cellphone and hotel reviews. According to our experiments, there are no obvious differences in opinion word extraction results for each of LRTBOOT and LSABOOT, given different numbers of seeds (e.g., one, five, ten, and so on), thus, we only showed their results using one single seed on the two review domains.

Table 3.6: Opinion word extraction performance on cellphone reviews.

Method	Seeds	Precision	Recall	F-measure
LRTBOOT	1	64.85%	76.04 %	70.00 %
LSABOOT	1	60.82%	71.17%	65.59%
DPHITS	1	62.66%	24.35%	35.08%
	5	65.75%	39.70%	49.51%
	10	66.27%	46.54%	54.68%
	15	66.10%	48.96%	56.25%
	20	65.92%	49.38%	56.46%
	25	65.98%	50.46%	57.18%
	50	65.74%	51.38%	57.78%
DP	1	60.94%	13.01%	21.44%
	5	64.46%	31.61%	42.42%
	10	64.98%	40.70%	50.05%
	15	64.17%	44.37%	52.47%
	20	63.78%	44.79%	52.62%
	25	63.58%	45.87%	53.30%
	50	63.21%	47.29%	54.10%

Cellphone Reviews

Table 3.6 lists the opinion word extraction results of all the evaluated methods on cellphone reviews. LRTBOOT achieves the best F-measure of 70.00%, which is 34.92% and 48.56% better than that of DPHITS and DP at single one seed, as well as 12.22% and 15.9% better than the two methods at all the fifty seeds. LSABOOT attains the second best F-measure of 65.59%, which is 30.51% and 44.15% larger than DPHITS and DP at single one seed, and also 7.81% and 11.49% better than both of them at the fifty seeds, respectively.

Owing much to the additional syntactic patterns and the HITS algorithm, DPHITS results in the overall best precision of 66.27% at the seed number of ten, but suffers from the poor coverage of the manually defined syntactic rules, leading to only 51.38% recall in its best case at the seed number of 50. Though giving comparable precision (its best precision is 64.98% at ten seeds), DP results in very low recall (47.29% at fifty seeds), and overall performs poorly compared to all other methods. The semantic dependency between pairwise words can be expressed subtly in customer reviews. It is generally difficult, if not impossible, for humans to recognize the relations to explicitly define a comprehensive set of syntactic dependency rules. As a result, a good many opinion words would be missed out incorrectly by the rule-based methods such as DP.

Thanks to the annotated domain-specific seeds, the opinion word extraction for each of DP and DPHITS has been improved a lot. Increasing the seed set size from one to twenty-five, DPHITS F-measure gains remarkably from 35.08% (1 seed) to 57.18(twenty-five seeds), while DP F-measure improves notably from 21.44% (one seed) to 53.30% (twenty-five seeds). Continuing to grow the seed number to fifty, DPHITS and DP are improved only a little to their best F-measure of 57.78% and 54.10%, respectively.

Hotel Reviews

Table 3.7: Opinion word extraction performance on hotel reviews.

Method	Seeds	Precision	Recall	F-measure
LRTBOOT	1	55.13 %	65.71%	59.96 %
LSABOOT	1	40.42%	62.27%	49.02%
DPHITS	1	50.62%	27.42%	35.56%
	5	54.23%	46.51%	50.08%
	10	52.92%	49.28%	51.04%
	15	52.76%	50.14%	51.42 %
	20	52.38%	52.53%	52.46%
	25	52.39%	53.39%	52.89%
	50	51.82%	54.54%	53.14%
DP	1	48.96%	22.45%	30.78%
	5	51.20%	44.70%	47.73%
	10	49.12%	47.95%	48.53%
	15	48.76%	48.90%	48.83%
	20	48.29%	51.39%	49.79%
	25	48.11%	52.25%	50.09%
	50	47.67%	53.77%	50.54%

Table 3.7 lists the opinion word extraction results on hotel reviews. LRTBOOT again achieves the overall best F-measure of 59.96%, which is 24.4% and 29.18% better than that of DPHITS and DP at the single one seed, and also 6.82% and 9.42% better than

both of them at the seed number of 50. LSABOOT attains the F-measure of 49.02%, which is 13.46% and 18.24% higher than that of DPHITS and DP at the single one seed, but 4.12% and 1.52% lower than the two methods at the 50 seeds. One illustration for this is that LSA model is less successful in discovering the hidden semantic structure of hotel review data, thus it may lead to undesirable effects on capturing the pairwise word dependencies when using the estimated corpus statistics associations based on the hidden structure. Pleasantly, LSABOOT results in much comparable recall of 62.27%, which is 7.73% and 8.5% better than DPHITS and DP at the fifty seeds.

The evaluation on hotel reviews again agrees with our findings that the proposed LRTBOOT on the whole results in the best opinion word extraction performance compared to the benchmark methods, and the seed set size has no significant effects on LRTBOOT as well as LSABOOT, but largely affects DP and DPHITS for opinion word extraction.

3.3.5 Discussion

Our proposed corpus statistics association based approach is largely language-independent, except for the candidate extraction step, where a linguistic tool is required to extract candidate aspect terms and opinion words from the corpus. As a first cut, the Stanford Parser can be used for English reviews to simply recognize nouns with subject or object syntactic pattern as candidate aspects. Likewise, the tagged adjectives or verbs can be recognized as the candidate opinion words in our approach. Once the candidates are extracted, our proposed methods, agnostic of the underlying language, will be used to discover domain-specific aspect terms and opinion words from customer reviews.

We tried to use opinion words as initial ground truth seeds. But the preliminary aspect and opinion extraction results deteriorate. One explanation may lie in that opinion words are not as domain-specific as aspect terms. A word is domain-specific, meaning that it is specific to the domain. The aspect terms of an entity are usually more indicative of the information related to the domain of the entity, while opinion words are primarily used to modify such domain-related aspect terms. For instance, in the cellphone reviews "The screen is amazing." and "It has an amazing camera.", the opinion word "amazing" is used as modifying opinion on the different aspects "screen" and "camera". The opinion word is also frequently used in the different hotel review domain, for instance, "The location is amazing." and "Amazing food!". Using opinion words as initial seeds could lead to many noisy aspects and opinions recognized. The extraction results then quickly deteriorate with additional bootstrapping iterations. The undesirable case can be considered as "blind leading the blind".

There are limitations in the proposed association-based bootstrapping method (ABOOT), which perhaps lead to the errors in the aspect and opinion extraction results.

(i) The proposed ABOOT method may suffer from the insufficient corpus statistics evidence problem for aspect and opinion extraction. For instance, the infrequent aspect terms like "repair rate" in cellphone reviews and "French windows" in hotel reviews would be missed out. Basically, This is not a serious problem, since the number of infrequent aspect terms is small by definition.

(ii) The non-noun aspect terms are missed out. For instance, valid verbal aspects, such as "video-recording" (cellphone domain) and "housekeeping" (hotel domain), cannot be recognized with ABOOT. This is partially because all non-noun words have not been considered when extracting candidate aspects.

(iii) Non-adjective or non-verb opinion words cannot be identified by the proposed ABOOT method. For instance, the noun "rubbish" is frequently used as a negative opinion word on cellphone domain, but it cannot be recognized via the method. There are also other noun opinion words that cannot be recognized on hotel domain, such as "hump" (aspect "bed") and "noise" (aspect "environment").

3.4 Summary

In this chapter, we propose a generalized corpus statistics association based bootstrapping approach (ABOOT) for aspect term and opinion word extraction. Starting with a small set of annotated aspect seeds, ABOOT can extract a large number of domain-specific aspect terms and the associated opinion words by mining semantic dependency patterns via corpus statistics association analysis.

Experimental results using real-world reviews demonstrate the benefits of the proposed method over benchmark methods. In fact, one instance method LRTBOOT (as well as LSABOOT) achieves very good performance for aspect term and opinion word extraction with only one aspect seed, which is simply the domain word, e.g., "cellphone" for cellphone reviews and "hotel" for hotel reviews. This makes our proposed association-based bootstrapping approach powerful and effective for practical aspect term and opinion word extraction in fine-grained opinion analysis of customer reviews.

In this chapter, we focus on the problem of extracting aspect terms and opinion words that explicitly appear in customer reviews. In the next chapter, we will study the implicit aspect identification problem, that is, identifying implicit aspects for the opinion words devoid of explicit aspects.

Chapter 4

Identifying Implicit Aspects via Cooccurrence Association Rule Mining

4.1 Introduction

In opinion mining, explicit aspect (term) extraction has been studied extensively, however, little work has been done on identifying implicit aspects, partially due to the challenging nature of the problem. In this chapter, we focus on the problem of implicit aspect identification, that is, identifying an implicit aspect for each recognized opinion word devoid of explicit aspect term in reviews. Implicit aspect refers to the aspect that does not appear but is indicated by an opinion word in a review.

Liu [Liu10] suggested to address the implicit aspect identification problem by mapping opinion words, also known as aspect indicators, to their underlying aspects based on a manually-compiled domain-dependent semantic dictionary. "However, it is not clear whether this is an effective approach as little research has been done" [Liu10].

Differently, we formulate the problem of inferring the implicit aspect for an opinion word devoid of explicit aspect term as a specialized form of association rule mining. The antecedent of an association rule corresponds to the opinion word, while the rule consequent indicates the aspect. An association rule is generated based on a cooccurrence matrix between a group of opinion words and a group of aspect terms.

We thus propose a two-phase cooccurrence association rule mining approach (coARM) for the implicit aspect identification problem. In the first phase, we generate a significant set of cooccurrence association rules by employing the measures support and confidence, where each rule associates an opinion word, i.e., rule antecedent, with an aspect term, i.e., rule consequent. We then group the rule consequents, i.e., aspect terms, into different aspect clusters to create more meaningful and robust association rules. In the second phase, for each new recognized opinion word devoid of explicit aspect term, we discover the matched robust rule whose consequent corresponds to the majority aspect cluster. Accordingly, the representative term of the aspect cluster is identified as the implicit aspect for the opinion word.

Experimental results show that coARM outperforms the well-established benchmark methods including a semantic dictionary based method, a statistical association method, and a mutual reinforcement clustering method.

We summarize our contributions in this work as follows:

- By formulating implicit aspect identification problem as a specialized form of association rule mining, we introduce a new two-phase cooccurrence association rule mining method (coARM) for the problem.
- To the best of our knowledge, this is the first work that conducts quantitative evaluations based on real-world reviews for implicit aspect identification.

4.2 Implicit Aspect Identification

4.2.1 Overview

The problem of finding the most likely aspect for an opinion word without explicit aspect can be formulated as a specialized type of association rule mining. An association rule is defined in the form [opinion-word → aspect-term], where the rule antecedent comes from an opinion word group, while the rule consequent is from another group of aspect terms. Each association rule is generated based on a cooccurrence matrix between the opinion word group and aspect term group. We thus call this approach cooccurrence association rule mining (coARM).

Essentially, the quality of cooccurrence association rule generation cane be assured using the significance measures support and confidence. This is because: 1) only the opinion words and aspect terms, whose support scores are larger than a pre-specified threshold, are considered to use as rule antecedents and rule consequents; 2) Moreover, only the association rules, whose confidence values exceed another pre-specified threshold, are considered as significant, while all other weak rules are filtered out.

In order to address the problem of implicit aspect identification, coARM first generates cooccurrence association rules from the reivew sentences which explicitly mention opinion words and the associated aspect term. Then, it applies the rules to an incomplete review sentence which typically mentions only opinion word but no the corresponding aspect term, thereby inferring the most likely implicit aspect. Figure 4.1 provides a simple example to illustrate how coARM works for implicit aspect identification.

In Figure 4.1, the proposed coARM method works in two phases of rule generation and rule application. In the first phase, a cooccurrence matrix is first built based on the complete review sentences which contains explicitly opinion words and aspect terms, where each entry indicates the cooccurrence frequency between an opinion word and an aspect (shown at the bottom). Then, given a minimum support 0.01 and a minimum confidence 0.3, a significant list of cooccurrence association rules are generated based on based on the cooccurrence matrix, where the antecedent of each rule indicates the opinion word group, while the rule consequent corresponds to the aspect term group. Further, for each rule antecedent, i.e., aspect term, the corresponding rule consequents,

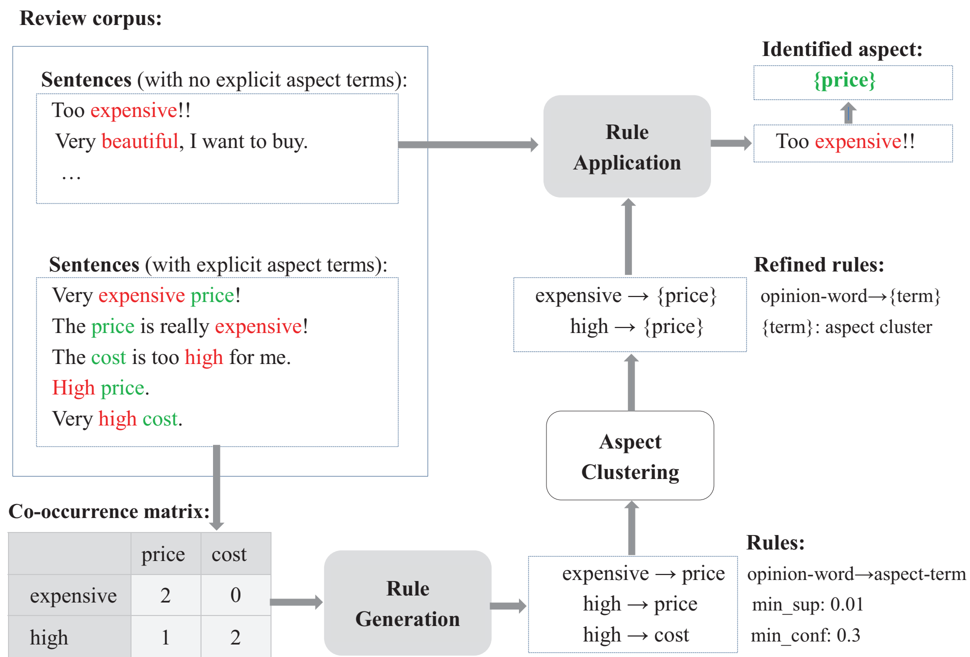

Figure 4.1: How coARM works for identifying implicit aspects.

i.e., aspect terms, are categorized into aspect clusters to generate much meaningful and robust rules. For instance, the two generated rules for the opinion word "high", [high $\rightarrow$ price] and [high $\rightarrow$ cost], have been grouped into one concise rule [high $\rightarrow$ {price}], where {price} means an aspect cluster that covers the aspects "price" and "cost".

In the second phase of rule application, for a new recognized opinion word "expensive", a refined cooccurrence association rule [expensive $\rightarrow$ {price}] is matched, and accordingly the representative term "price" of its consequent aspect cluster is identified as the implicit aspect for the opinion word.

4.2.2 Cooccurrence Association Rule Mining

In this section, we describe our new cooccurrence association rule mining method in detail.

Building Opinion Word Group and Aspect Term Group

In the first phase of coARM, the association rule generation replies on two groups of opinion words and aspect terms. In this work, we simply used a candidate group of opinion words and a candidate group of aspect terms instead. This is because noisy opinion words, aspect terms, as well as the unlikely dependency between them can be largely pruned based on the significance measures support and confidence.

Typically, aspect terms often occur as nouns or noun phrases with certain syntactic patterns in sentences, such as *subject-verb* ("SBV"), *verb-object* ("VOB"), *preposition-object* ("POB"), or even *head word* ("HED"). Thus to form the candidate group of aspect terms G_A, we extract all the nouns or noun phrases with the "SBV", "VOB", "POB", or "HED" syntactic relationships in individual sentences of the review corpus.

In addition, opinion words are likely to appear as adjectives or verbs in sentences. Then, to form the candidate group of opinion words G_O, we recognize all the adjectives and verbs, which appear together with the extracted candidate aspect terms in individual sentences. Note that the adjectives or verbs with the syntactic dependency relation of *adverbial* ("ADV") are not considered, this is because the words with pattern "ADV" tend to appear as modifiers of true opinion words.

Building Cooccurrence Matrix

We build a cooccurrence matrix M_{OA} between the opinion word group G_O and the aspect term group G_A. Each row of the matrix M_{OA} corresponds to an opinion word o_i, while each column corresponds to an aspect term a_j. The entry of the matrix represents the number of times an opinion word and an aspect term cooccur in a piece of text.

The cooccurrence matrix can be constructed at various resolutions, i.e., sentence, paragraph, or document. We find experimentally that the sentence resolution leads to much more meaningful semantic associations.

Formalizing Cooccurrence Association Rule Mining

Let D denote a set of transactions. Let itemset $I_1 = \{i_{11}, i_{12}, \ldots, i_{1M}\}$ denote the set of M items with the label L_1, and itemset $I_2 = \{i_{21}, i_{22}, \ldots, i_{2N}\}$ denote the set of N items with the label L_2, where $I_1 \cap I_2 = \varnothing$, Let itemset $I = I_1 + I_2$ be the complete set of items. Let $M_{L_1 L_2}$ be an $M \times N$ cooccurrence matrix, where each entry m_{ij} represents the cooccurrence frequency between a pair of items. A cooccurrence association rule is defined in the form:

$$X \to Y, \tag{4.1}$$

where $X \in I_1$ and $Y \in I_2$ are the items from the sets I_1 and I_2, respectively.

The rule $X \to Y$ has the support score s in D if s of all transactions in the data set D contain both X and Y, while the rule holds in D with confidence value c if c of transactions in D that contain X also contain Y.

Generating Cooccurrence Association Rules

Next, we apply the cooccurrence association rule mining to a given review domain to generate association rules.

Specifically, let $D = \{s_1, s_2, \ldots, s_R\}$ denote a review corpus containing R review sentences s_r, with $r = 1, \ldots, R$. $G_O = \{o_1, o_2, \ldots, o_M\}$ indicate a group of opinion word o_i, corresponding to the aforementioned set I_1. $G_A = \{a_1, a_2, \ldots, a_N\}$ indicate a group of aspect term a_j, corresponding to the itemset set I_2. $M_{OA} = (m_{ij})_{M \times N}$ indicate a cooccurrence matrix, where each entry m_{ij} represents a cooccurrence sentence frequency between the opinion word o_i and the aspect term a_j.

For each opinion word o_i in the opinion group G_O, we initially generate a list of association rules with every aspect term a_j in the aspect group G_A based on the cooccurrence matrix M_{OA} in the form:

$$o_i \to a_j, \tag{4.2}$$

Typically, each association rule has a support score sup_{ij} calculated as follows:

$$sup_{ij} = \frac{\#(o_i \cup a_j)}{R}, \tag{4.3}$$

where $\#(o_i \cup a_j)$ is the cooccurrences sentence frequency between opinion word o_i and aspect term a_j, and R is the total number of review sentences.

The confidence score $conf_{ij}$ is calculated as follows:

$$conf_{ij} = \frac{\#(o_i \cup a_j)}{\#(o_i)}, \tag{4.4}$$

where $\#(o_i)$ is the number of review sentences containing o_i.

Given appropriate support and confidence thresholds min_sup and min_conf, we prune the generated list of cooccurrence association rules by discarding the rules whose support and confidence scores are less than the thresholds. We generate only significant association rules for the application to implicit aspect identification. Algorithm 2 summarizes the cooccurrence association rule mining process.

Algorithm 2 Cooccurrence association rule mining.

Require: Review corpus $\mathcal{D}$
Ensure: Generated cooccurrence association rules
$G_O \leftarrow$ Extract an opinion word group from corpus $\mathcal{D}$;
$G_A \leftarrow$ Extract an aspect term group from corpus $\mathcal{D}$;
$M_{OA} \leftarrow$ Build a cooccurrence matrix between groups G_O and G_A;
for *each opinion word* o_i *in the group* G_O **do**
  for *each aspect term* a_j *in the group* G_A **do**
    Form a cooccurrence association rule: $o_i \rightarrow a_j$;
    Compute support and confidence of the rule via the equations 4.3 and 4.4;
    if $(sup_{ij} \geq min\_sup) AND (conf_{ij} \geq min\_conf)$ **then**
      Generate a validated cooccurrence association rule in set S;
    end if
  end for
end for
return A validated set of cooccurrence association rules S

Clustering Aspect Terms

Generally, people often use different aspect terms to represent the same product aspect in practice. It is thus more meaningful and concise to associate an opinion word to the most likely aspect cluster instead of a single aspect term. An aspect cluster categorizes a set of synonymous aspect terms. For instance, the aspect cluster of “appearance” categorizes a list of aspect terms like “appearance” and “exterior” on cellphone review doamin.

Thus, for each opinion word, also known as rule antecedent, the generated cooccurrence association rules are refined by clustering the rule consequents, i.e., aspect terms. After that, the new consequent of each refined rule is no long a specific aspect term but an aspect cluster, which could be more meaningful in practice. For instance, we first generate a list of cooccurrence association rules for the opinion word “beautiful” on the cellphone reviews, such as [beautiful $\rightarrow$ appearance] and [beautiful $\rightarrow$ exterior]. We then refine the generated rules for the opinion word as [beautiful $\rightarrow$ {appearance}], where {appearance} represents the aspect cluster covering the particular aspect terms “appearance” and “exterior”.

We use the well-known *k-means* algorithm to cluster the aspect terms. Each aspect term is represented by using the contextual word features (nouns, verbs, adjectives, and adverbs) that cooccur with the aspect term in the same review sentences. We evaluated different values of the aspect cluster number and specified K=*22* clusters on the cellphone review corpus. In addition, each aspect cluster is indicated by its representative aspect term which has the largest corpus frequency within that cluster.

4.2.3 Implicit Aspect Identification via coARM

With the refined cooccurrence association rules, we can proceed to the next phase of applying the rules for implicit aspect Identification.

First of all, we need to extract a new set of opinion words devoid of explicit aspect terms from a given review corpus, which is a challenging problem. In this work, we simply locate the adjectives or verbs without the syntactic pattern "ADV" from the incomplete review sentences, which typically mention opinion words but no the corresponding aspect terms.

Next, for each new recognized opinion word (devoid of explicit aspect), we match one refined cooccurrence association rule whose consequent indicates the majority aspect cluster. Then, we identify the representative term of the aspect cluster as the implicit aspect for the opinion word.

For example, an opinion word "beautiful" is recognized from this cellphone review "It is really beautiful, I will get one.", but the corresponding opinionated aspect is not mentioned. With applying the proposed coARM approach, We can match a cooccurrence association rule to the opinion word, [beautiful $\rightarrow$ {appearance}], where $\{appearance\}$ means an aspect cluster indicated by the representative aspect term "appearance". Then, we assign the representative term "appearance" to the opinion word "beautiful" as the inferred implicit aspect.

In reality, it is possible that a new recognized opinion word does not fire any cooccurrence association rules. In this case, we can rely on an external synonym/antonym dictionary such as WordNet [MBF$^+$90]. That is, if a synonym or antonym of the recognized opinion word matches a refined rule, then the representative term of the rule consequent (i.e., aspect cluster) will be inferred as the implicit aspect for the opinion word.

All in all, the proposed cooccurrence association rule mining approach (coARM) exploits the frequent explicit interdependent evidence on the review corpus to derive the most likely aspect for each opinion word devoid of explicit aspect in a review sentence. The coARM approach is highly adaptable to any domain or corpus size, although a larger corpus tends to naturally give better results.

4.3 Experiments

We evaluated the implicit aspect identification of the proposed coARM method on real-world product review corpus. The corpus contains 2,986 cellphone reviews collected from a Chinese website [1].

To create golden standard for evaluation, we first randomly selected 351 review documents, and also provided a list of candidate implicit aspect terms on the domain. For each review sentence in the selected reviews, two persons manually annotated the opinion words devoid of explicit aspect terms, and recognized the matched implicit aspects from the provided list. We finally collected 449 opinion word and implicit aspect pairs in the golden standard set.

4.3.1 Experimental Setup

We conducted two types of experiments, one is to evaluate the implicit aspect identification of coARM against several well-established baseline methods, the other is to evaluate the performance of coARM versus confidence scores. We described the baselines below.

To identify implicit aspects for opinion words devoid of explicit aspects, semantic dictionary based method, named Dictionary, simply maps the opinion words to the frequently associated aspect terms by using a semantic dictionary, which contains pre-specified opinion-aspect pairs [HCSK10]. A likelihood ratio tests [Dun93] based method (LRT) identifies the implicit aspects for the opinion words via statistical association analysis on the given review corpus. A mutual reinforcement clustering method (MRC) [SXG+08] infers the implicit aspects by using a hidden semantic association set between opinion word categories and aspect clusters, which is constructed by employing mutual reinforcement clustering principle.

[1] product.mobile.163.com

4.3.2 Implicit Aspect Identification Results

We first evaluated the implicit aspect identification of coARM against several benchmark methods, namely, Dictionary, LRT, and MRC. Table 4.1 shows thee best results for all evaluated methods.

Table 4.1: Implicit aspect identification results of different methods.

Method	Precision	Recall	F-measure
Dictionary	58.44%	58.84%	58.64%
LRT	59.62%	56.82%	58.19%
MRC	79.56%	56.60%	66.14%
coARM	76.29%	72.71%	74.46%

The proposed coARM method achieves the best F-measure of 74.46%, which is 15.82%, 16.27%, and 8.32% better than that of the baseline methods Dictionary, LRT, and MRC, respectively.

The Dictionary method tends to suffer from the low-quality opinion-aspect pairs collected in the semantic dictionary, leading to poor performance (58.64% F-measure). LRT is also prone to the noisy corpus statistics associations between opinion words and aspect terms.

MRC achieves 66.14% F-measure, 79.56% precision, and 56.60% recall. MRC is much effective for implicit aspect identification compared to the baselines Dictionary and LRT. However, its F-measure is still 8% lower than coARM. MRC's precision is the highest among all methods, while its recall is 16.11% worse than coARM. Basically, MRC tends to suffer from the presence of erroneous mutual dependency relations between candidate opinion words and aspect terms during clustering. In addition, MRC's low recall could be also attributed to the fact that it only considers adjectives as opinion words while coARM considers both adjectives and verbs. However, adding more candidate verbal opinion words into the MRC framework makes clustering even harder, leading to worse performance according to our experiments.

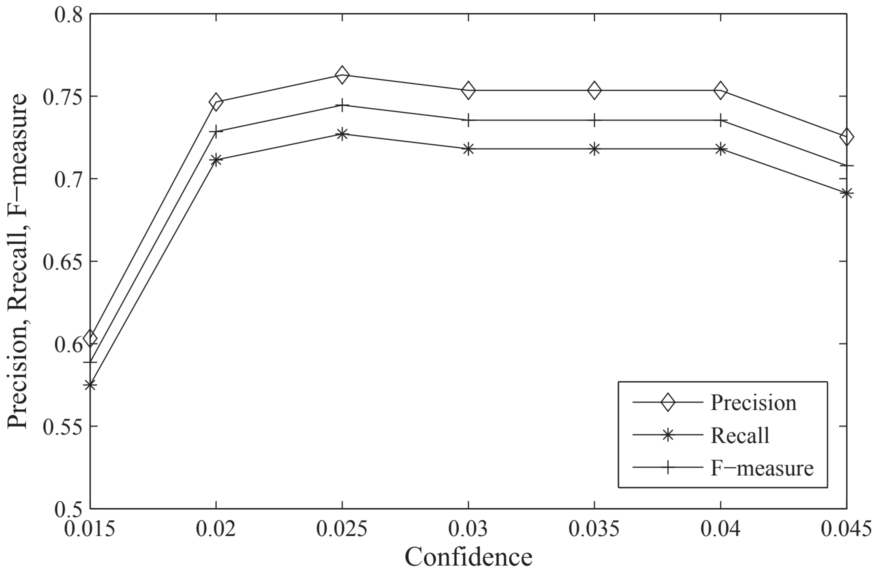

Figure 4.2: Performance vs. confidence threshold for coARM implicit aspect identification.

Figure 4.2 plots coARM performance versus different confidence thresholds, while keeping the support threshold fixed at 0.0005. The curves of the F-measure, precision, and recall of coARM initially go up when increasing confidence values, achieving the best F-measure of 74.46% at 0.025 confidence threshold. If we continued to increase the confidence threshold, and the performance curves go down. This is expected as a higher confidence threshold will weed out the noisy insignificant association rules used for implicit aspect inference, but increasing the confidence threshold beyond a certain point will also remove some meaningful association rules as well.

4.4 Summary

In this chapter, we propose a two-phase cooccurrence association rule mining method (coARM) for identifying an implicit aspect for each opinion word devoid of explicit aspect in reviews. The experimental results using real-life reviews demonstrate that coARM outperforms the well-established benchmark methods.

In Chapter 3, we focus on extracting specific aspect terms and opinion words that appear in customer reviews. One limitation of the work lies in that there is redundancy in the review mining and summarization results, this is because different people often use different aspect terms to express the same semantic aspect. For example, the aspect terms "screen", "LCD", and "display" are all indicative of the single aspect "screen" in cellphone review domain. A separate step of categorization may be used to group the synonymous aspect terms. In this chapter, we have used k-means algorithm to cluster the rule consequents, i.e., aspect terms, to generate robust cooccurrence association rules. However, a separate step of clustering like this may lead to the accumulation of errors.

Instead, in the next chapter, we develop an advanced supervised joint topic model to identify semantic aspects and opinions on the aspects and to predict overall review ratings in a unified framework.

Chapter 5

Modeling and Mining Reviews via Supervised Joint Topic Model

5.1 Introduction

In this chapter, we focus on the problem of modeling customer reviews with aims at identifying semantic aspects and aspect-based opinions from reviews as well as predicting overall review ratings in a unified framework.

Recently, probabilistic topic models have been shown effective for mining semantic topical structure of textual data. Different extensions based on a basic latent Dirichlet allocation model (LDA) [BNJ03] have been developed to address opinion analysis problem [TM08b, LH09, ME11]. As far as we know, most of existing topic modeling approaches are unsupervised. One limitation of these models is that detecting of latent aspects/topics and opinion orientations from reviews is blind.

Differently, we propose to incorporate observed overall ratings of customer reviews into a unified topic modeling framework as supervision knowledge. We also reduce each review document as a bag of opinion pairs, where each pair contains an aspect term and its associate opinion word. We introduce a supervised joint aspect and opinion model (SJAOM), a generative probabilistic supervised joint topic model for fine-grained opinion analysis of customer reviews. SJAOM can simultaneously model the aspect terms and opinion words in review texts, and jointly infer aspects and aspect-based opinions under the supervision of the overall rating information. One key benefit of SJAOM is that it can discover the semantic aspects and opinions that are predictive of the overall ratings of reviews.

We extensively validate SJAOM for opinion analysis problem using publicly available review data from three categories, and experimentally demonstrate the superiority of SJAOM against six existing representative baseline methods, in particular, a supervised latent Dirichlet allocation model (sLDA) [BM07], three weakly/partially supervised topic models called joint sentiment-topic model (JST) [LHER12], aspect-sentiment unification model (ASUM) [JO11], and latent aspect rating analysis model (LARA) [WLZ10], and two non-generative models, namely, a supervised support vector machines sentiment classifier (SVM) [PLV02] and an unsupervised lexicon-based method (Lexicon) [TBT+11].

We summarize our contributions as follows:

- We propose a novel supervised joint topic model called SJAOM. One key advantage of SJAOM is that it can jointly discover the hidden aspects and opinions that are indicative of overall ratings of reviews.

- To the best of our knowledge, our approach is the first to use supervised joint topic models to address opinion mining and analysis problems.

- Given a new unlabeled review, we form the prediction of overall rating via a carefully-designed regression procedure on the detected aspects and opinions in the review. This is different from existing topic modeling approaches [JO11, LHER12], which typically formulate the overall opinion prediction as a classification problem and simply classify each review based on the per-document opinion distributions.

- Experimental results using publicly available review data illustrate the benefits of SJAOM for opinion analysis tasks against six existing representative methods.

5.2 Supervised Joint Aspect and Opinion Model of Mining Reviews

5.2.1 Problem Definition

We focus on modeling customer reviews with objectives of identifying semantic aspects and opinion orientations on the aspects from reviews as well as predicting overall review ratings in a unified framework.

We define our problem to be addressed in this work in detail as following:
Given a product from a category, there is a collection of M customer reviews (documents) on the product: $D = \{d_1, d_2, \ldots, d_M\}$. Each review d_m is reduced to a set of N_m opinion pairs: $d_m = \{\langle t_1, w_1\rangle, \langle t_2, w_2\rangle, \ldots, \langle t_{N_m}, w_{N_m}\rangle\}$, where each opinion pair consists of an aspect term t_n and the corresponding opinion word w_n in the review. Our task is to identify K semantic aspects of the product and opinion orientations on the aspects, as well as to predict the overall ratings of unlabeled reviews.

In particular, we aim to address three opinion analysis tasks:

- Aspect detection. We focus on inferring the semantic aspects of the reviewed product by grouping or clustering the synonymous or semantically related aspect terms and opinion words.

- Aspect-based opinion identification. We identify which opinion and sentiment orientations are expressed towards the detected aspects in reviews.

- Overall rating prediction. Given a new unlabeled review, we aim to form the prediction of its overall rating via a fitted SJAOM model based on the detected latent aspects and opinion orientations in the review.

5.2.2 Overview

Probabilistic topic models have been shown effective for mining hidden topical structure of textual data [BNJ03, GS04a]. We employ topic modeling technique to solve the aspect-based opinion analysis problem. In addition to the module of detecting semantic aspects from reviews, we also need to develop a counterpart for identifying opinions on the aspects.

Different extensions based on the basic latent Dirichlet allocation model (LDA) [BNJ03] have been proposed for aspect-based opinion analysis problem. Titov and McDonald [TM08a] introduced an unsupervised multi-grain LDA model (MG-LDA) for extracting semantic aspects commented on in reviews. They then incorporated a latent aspect rating analysis module in MG-LDA, and introduced a multi-aspect sentiment model (MAS) for aspect-based sentiment summarization [TM08b]. Lin and He [LH09] extended the standard LDA by adding a sentiment layer and proposed an a joint sentiment and topic model (JST) for sentiment analysis of reviews. Moghaddam and Ester [ME11] proposed an interdependent LDA model (ILDA) for mining latent aspects and the aspect-dependent sentiments from textual reviews. Most of existing topic modeling methods for opinion analysis are unsupervised. According to our experiments, without the supervision knowledge, such as overall ratings of reviews, the unsupervised detection of latent aspects or opinions tends to be blind, and the aspect-based opinion summarization results are also hard to interpret.

Generally, customer reviews often come with overall ratings, for instance, in the form of one-to-five stars, which are valuable and insightful for discovering customer reviews. We propose to model the overall ratings as guidance knowledge to supervise the process of identifying semantic aspects and aspect-specific opinions from reviews. In addition, the vast majority of existing probabilistic topic modeling approaches employ the bag-of-words representation of review documents. Differently, we reduce each review as a bag of opinion pairs, where each opinion pair consists of an aspect term and the corresponding opinion word.

We introduce a supervised joint topic model named supervised joint aspect and opinion model (SJAOM) for opinion analysis of customer reviews. SJAOM incorporates the overall review ratings as supervision data, and simultaneously models the pairwise aspect terms and opinion words in a unified framework. One key advantage of SJAOM is its ability to jointly identify the semantic aspects and opinions that are predictive of the overall ratings of reviews.

5.2.3 Supervised Joint Aspect and Opinion Model

We make the following assumptions for the generative process of our proposed SJAOM:

- The generation of opinion orientations depends on semantic aspects, meaning that we first generate the aspects commented on in reviews, on which we subsequently generate the associated opinion orientations. Intuitively, different product aspects often have different utility qualities, which thus lead to different evaluations and sentiments.

- Aspect term generation depends on the aspects, and opinion word generation relies on the opinion orientations as well as the associated semantic aspects. For instance, to generate an opinion word "beautiful", we need to generate not only the opinion orientation "positive" but also the related hidden aspect "appearance".

- The generation of overall ratings are based on the hidden aspects and opinion orientations generated in individual reviews.

Based on the assumptions, to generate a review document and its real-valued overall rating by SJAOM, aspects are first generated conditioned on the per document aspect distribution. The opinion orientation on an aspect is then generated conditioned on per document aspect-specific opinion distribution. Each opinion pair contains an aspect term and an opinion word, and it is generated conditioned on the respective word distributions. The overall rating is finally generated based on all of the aspect and opinion assignments in the review.

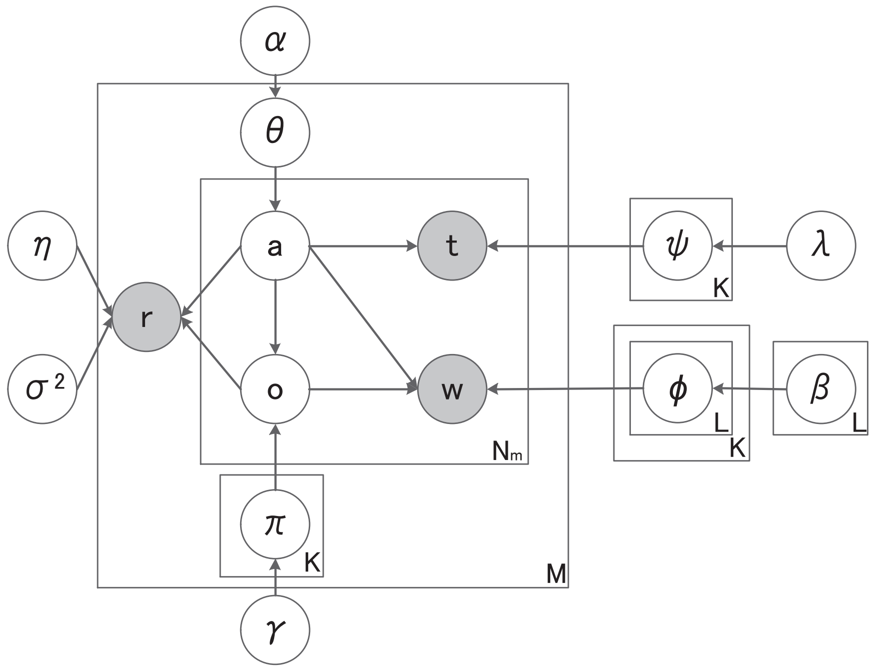

Figure 5.1: Graphical model representation of SJAOM.

The graphical model representation of SJAOM is shown in Figure 5.1, and the notations in this model are listed in Table 5.1. Under SJAOM, each review document d_m and its real-valued overall rating r_m are generated from the following processes:

- For each aspect $k \in \{1, \ldots, K\}$
  - (i) Draw aspect word distribution $\psi_k \sim \mathrm{Dir}(\lambda)$.
  - (ii) For each opinion orientation $l \in \{1, \ldots, L\}$
    - (a) Draw opinion word distribution $\phi_{kl} \sim \mathrm{Dir}(\beta_l)$.
- For each review d_m and its overall rating r_m

Table 5.1: Notations used in SJAOM.

M	Number of review documents in a corpus
N_m	Number of opinion pairs in review d_m
K	Number of semantic aspects
L	Number of opinion orientations
r_m	Overall rating of review document m
t_{mn}	Aspect term in the nth opinion pair of review d_m
w_{mn}	Opinion word in the nth opinion pair of review d_m
a_{mn}	Aspect assignment to aspect term t_{mn} and opinion word w_{mn}
o_{mn}	Opinion orientation assignment to opinion word w_{mn}
θ	Dirichlet distribution over aspects
π	Dirichlet distribution over opinion orientations
ψ	Dirichlet distribution over aspect words
ϕ	Dirichlet distribution over opinion words
α	Hyperparameter for aspect distribution θ
γ	Hyperparameter for opinion distribution π
λ	Hyperparameter for aspect word distribution ψ
β	Hyperparameter for opinion word distribution ϕ
η	Rating response parameters
σ^2	Rating response parameters
U	Number of unique aspect words in vocabulary
V	Number of unique opinion words in vocabulary
$\boldsymbol{a}^{\neg i}$	All aspect assignments except for a_i
$\boldsymbol{o}^{\neg i}$	All opinion orientation assignments except for o_i
$N_{m,k}$	Count of words in document m assigned to aspect k
$N_{m,k,l}$	Count of words in document m assigned to aspect k and opinion orientation l
$N_{k,u}$	Count of unique aspect word u in vocabulary assigned to aspect k
N_k	Total count of aspect words in vocabulary assigned to aspect k.
$N_{k,l,v}$	Count of unique opinion word v in vocabulary assigned to aspect k and opinion l
$N_{k,l}$	Total number of opinion words in vocabulary assigned to aspect k and opinion l

(i) Draw aspect distribution $\theta_m \sim \text{Dir}(\alpha)$.

(ii) For each aspect k under review m

  (a) Draw opinion distribution $\pi_{mk} \sim \text{Dir}(\gamma)$.

(iii) For each opinion pair $\langle t_{mn}, w_{mn} \rangle, n \in \{1, \ldots, N_m\}$

  (a) Draw aspect assignment $a_{mn} \sim \text{Mult}(\theta_m)$.

  (b) Draw opinion orientation assignment $o_{mn} \sim \text{Mult}(\pi_{ma_{mn}})$.

  (c) Draw aspect term $t_{mn} \sim \text{Mult}(\psi_{a_{mn}})$.

(d) Draw opinion word $w_{mn} \sim \mathrm{Mult}(\phi_{a_{mn} o_{mn}})$.

(iv) Draw review rating response $r_m \sim \mathrm{N}(\eta^T \bar{z}_m, \sigma^2)$, where we define

$$\bar{z}_m = \frac{1}{C} \sum_{n=1}^{N_m} (a_{mn} \times (\omega^T \times o_{mn})). \tag{5.1}$$

The quantity $\bar{z}_m$ represents the combined empirical frequencies of the latent aspects and opinions in a review document r_m, or simply the normalized empirical frequencies of the sentimental (e.g., positive or negative) aspects in the review, where C is a normalization constant, and ω represents the weights for the hidden opinion orientations, which can be estimated experimentally via grid-search on the training data.

The real-valued overall rating response r_m is drawn from a normal linear model. The normalized empirical frequencies $\bar{z}_m$ works as the covariates in this model, η indicates the regression coefficients on the frequencies. The parameters $\eta^T \bar{z}_m$ and σ^2 are the mean and variance of the normal distribution.

We regress the overall rating response of a review on the normalized empirical frequencies of the underlying positive and negative aspects in the review, meaning that the aspect terms, opinion words, as well as their aspect and opinion orientation assignments in the review are generated first, then, based on the generated hidden aspects and opinions of the review, the overall rating response is finally generated.

The formulation agrees with our intuitions that: 1) Generally, different products have particular sets of aspects with diverse qualities, which lead to different experience, evaluations, and sentiments on the aspects, and then 2) Overall ratings of reviews on a product are formed based on the individual sentimental aspects of the product in the reviews, where the regression coefficients η capture the relative contributions of the latent positive or negative aspects on generating the overall rating. By doing the fine-grained opinion analysis, we can understand why a product receives a particular overall rating in a review (e,g,. one to five stars), and which positive or negative aspects contribute to the overall rating.

In addition, we note that one limitation of existing unsupervised topic models is that the correspondence between detected latent opinion variables/clusters and the actual sentimental orientations is not explicit [TM08b, ME11]. To explicitly build the connection, SJAOM exploits sentiment prior information by using asymmetric hyperparameter β for opinion word distribution based on a opinion dictionary in the modeling framework.

5.3 Inference and Prediction

In this section, we describe the approximate inference and parameter estimation for the proposed SJAOM. We also describe how to apply SJAOM to opinion mining and analysis problem.

5.3.1 Inference and Parameter Estimation

Our target of inference is to evaluate the posterior distribution of the hidden variables given a review document and its rating:

$$p(\boldsymbol{a}, \boldsymbol{o} \mid \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2) = \frac{p(\boldsymbol{a}, \boldsymbol{o}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2)}{p(\boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2)}. \tag{5.2}$$

The exact inference for the posterior distribution is intractable. We use a collapsed Gibbs sampling algorithm [GG84, GS04b] for the approximate inference of SJAOM.

For an opinion pair $\langle t_{mn}, w_{mn} \rangle$ with the index $i = (m, n)$, we compute the full conditional distribution as following:

$$\begin{aligned} & p(a_i = k, o_i = l \mid \boldsymbol{a}^{\neg i}, \boldsymbol{o}^{\neg i}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2) \\ & \propto \frac{N_{m,k}^{\neg i} + \alpha_k}{N_m^{\neg i} + \sum_{k'} \alpha_{k'}} \cdot \frac{N_{m,k,l}^{\neg i} + \gamma}{N_{m,k}^{\neg i} + L\gamma} \cdot \frac{N_{k,u}^{\neg i} + \lambda}{N_k^{\neg i} + U\lambda} \cdot \\ & \frac{N_{k,l,v}^{\neg i} + \beta_{l,v}}{N_{k,l}^{\neg i} + \sum_{v'} \beta_{l,v'}} \cdot \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(r_m - \eta^T \bar{z}_m^{\neg i})^2}{2\sigma^2}). \end{aligned} \tag{5.3}$$

Notations used in Equation 5.3 are listed in in Table 5.1. The detailed derivation of this equation is given in Appendix A.

The quantities $\boldsymbol{a}^{\neg i}$ and $\boldsymbol{o}^{\neg i}$ indicate the assignments of aspects and opinion orientations to all aspect terms and opinion words in the corpus except for the assignments a_i and o_i for the aspect term and opinion word at position i, respectively. The $N_{m,k}$ is the count of aspect k assigned to words in review document m, and the $N_{m,k,l}$ means the count of opinion orientation l allocated to words that are assigned to aspect k in review m. The quantity $N_{k,u}$ indicate the number of unique aspect word u in vocabulary assigned to aspect k, and the $N_{k,l,v}$ is the number of unique opinion word v in vocabulary allocated to opinion orientation l and aspect k across all documents. The subscript $-i$ in a quantity indicates the exclusion of the corresponding data at index i. For example, this quantity $N^{\neg i}_{k,l,v}$ indicates the count of the vocabulary opinion word v assigned to aspect k and opinion l, excluding the opinion word and the corresponding opinion orientation assignment at index i.

Next, we compute the SJAOM model parameters based on the samples collected via the Gibbs sampling algorithm as follows. We compute the per document aspect distribution:

$$\theta_{m,k} = \frac{N_{m,k} + \alpha_k}{N_m + \sum_{k'=1}^{K} \alpha_{k'}}. \tag{5.4}$$

We compute the per document aspect specific opinion distribution:

$$\pi_{m,k,l} = \frac{N_{m,k,l} + \gamma}{N_{m,k} + L\gamma}. \tag{5.5}$$

We compute the aspect word distribution:

$$\psi_{k,u} = \frac{N_{k,u} + \lambda}{N_k + U\lambda}. \tag{5.6}$$

We compute the opinion word distribution as follows:

$$\phi_{klv} = \frac{N_{k,l,v} + \beta_{l,v}}{N_{k,l} + \sum_{v'=1}^{V} \beta_{l,v'}}. \tag{5.7}$$

We use asymmetric Dirichlet prior α for the per document aspect distribution θ, and estimate the hyperparameter α by employing a fixed-point iteration scheme [Min00]. To explicitly build the correspondence between latent opinion variables and real sentimental orientations, SJAOM exploits prior information by using asymmetric hyperparameter β for opinion word distribution ϕ. The opinion prior knowledge comes from a publicly available lexicon called MPQA Subjectivity Lexicon [1]. In our experiments, we consider two opinion orientations positive and negative ($L = 2$). The elements of β corresponding to positive opinion words like "great" and "perfect" have very high values for positive orientation, namely, β_{lv} =0.95, and very low values for negative orientation, namely, β_{lv} =0.05. The *beta* values for negative opinion words like "bad" and "poor" are assigned conversely. Following previous work [GS04b, LHER12], We use symmetric priors γ and λ for per document aspect specific opinion distribution π and aspect word distribution ψ, and simply set them as $1/L$ and 0.01, respectively.

We follow Blei and McAuliffe [BM07] to approximately evaluate the normal linear model parameters η and σ^2. Let Z be the $M \times L$ matrix whose rows are the vectors $\bar{z}_m^T$. Then, the regression coefficients η is approximated as follows:

$$\hat{\eta} \approx (Z^T Z)^{-1} Z^T r, \tag{5.8}$$

where the vector r consists of the overall ratings of all reviews.

We estimate σ^2 as follows:

$$\hat{\sigma}^2 \approx \frac{1}{M}\{r^T r - r^T Z (Z^T Z)^{-1} Z^T r\}. \tag{5.9}$$

5.3.2 Overall Rating Prediction

Our SJAOM model formulates the overall rating prediction of a review as a regression problem. In particular, given a new unlabeled review $d_{m'}$ and a fitted SJAOM model

[1] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon

$\{\psi_k, \phi_{kl}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2\}$ $(k : 1 \sim K, l : 1 \sim L)$, our idea for the overall rating prediction is to learn the latent aspects and opinions over the opinion pairs in the review, and then approximately form the regression function on the posterior mean $\bar{z}_{m'}$, as shown below:

$$\hat{r}_{m'} \approx \eta^T \bar{z}_{m'}. \tag{5.10}$$

Similarly, we approximate the posterior mean of $\bar{z}_{m'}$ by applying Gibbs sampling as described in the previous section. Note that the parts relying on the overall rating variable are removed from the sampling equation, as new testing review documents contain no rating annotations.

5.4 Experiments

In this section, we describe and discuss the experimental results of opinion analysis on real-world customer review data.

5.4.1 Experimental Setup

We conducted three experiments to evaluate the proposed SJAOM model as follows:

- Identifying semantic aspects. We detected K aspects of a reviewed entity (e.g., product) via SJAOM, and evaluated the aspect detection using Rand index [Ran71], a measure of the similarity between two data partitions. The value of Rand index lies between zero and one, where zero means that the two data partitions do not agree on any pair of elements, while one means the partitions are exactly the same.

- Identifying opinions on aspects. We identified the opinion orientations expressed towards the particular semantic aspects, and evaluated the aspect-specific opinion identification via accuracy measure.

- Predicting overall opinion ratings. Given a new unlabeled review, we formed the prediction of the overall opinion rating based on the detected latent aspects and opinions via a fitted SJAOM model. We used an accuracy measure to evaluate the performance.

We experimentally compared SJAOM against six existing benchmark models: four generative models called supervised latent Dirichlet allocation model (sLDA) [BM07], joint sentiment-topic model (JST) [LHER12], aspect-sentiment unification model (ASUM) [JO11], and latent aspect rating analysis model (LARA) [WLZ10], and two non-generative models, namely, a supervised SVM sentiment classification method (SVM) [PLV02] and a unsupervised lexicon-based method (Lexicon) [TBT+11].

Note that the generative models SJAOM, sLDA, JST, and ASUM were evaluated for all the aforementioned tasks. The LARA model was evaluated in both tasks of aspect detection and aspect-specific opinion identification. This is because LARA is developed for latent aspect rating analysis and cannot be used for prediction of the overall opinion rating of a customer review. The last two non-generative baselines SVM and Lexicon were used only in the evaluation of overall opinion prediction, as the two methods cannot model the latent topical structure of review data.

In our experiments, we used the publicly available implementations for the baseline methods sLDA [2], JST [3], ASUM [4], LARA [5], and SVM^{light} package [6].

5.4.2 Data Sets

We evaluated SJAOM using publicly available customer review data from three categories, namely, game (video game), CD (audio CD), and hotel. Game and CD review

[2] www.cs.cmu.edu/chongw/slda
[3] sites.google.com/site/yulanhe8/code
[4] uilab.kaist.ac.kr/research/WSDM11
[5] sifaka.cs.uiuc.edu/wang296/Codes/LARA.zip
[6] svmlight.joachims.org

data sets [7] are collected from Amazon which is the world's largest online retailer [8], and hotel review data set [9] is collected from TripAdvisor which is the largest social travel community in the world [10]. Some statistics of the data sets are listed in Table 5.2.

Table 5.2: Some statistics of review data sets.

Category	Game	CD	Hotel
# of reviews in corpus	2599	1632	1,367
# of words in corpus	554,496	292,060	696,327
Vocabulary size	23,809	13,886	21,785
Average # of words/review	213	178	509

In preprocessing the data, we parsed all the reviews in each data set using the Stanford Parser [KM03]. We applied the three grammatical dependency relations, *adjectival modifier* ("amod"), *direct object* ("dobj"), and *nominal subject* ("nsubj") to opinion pair extraction from each parsed review. We then extended the extracted set of opinion pairs via applying two additional dependency relations *negation modifier* ("neg") and *conjunct* ("conj").

For each of the three data sets, we held out 20% of the data for testing purpose and trained all the evaluated models on the remaining 80% of the data.

Based on each given testing data, a list of 150 frequent domain-dependent keywords, including aspect terms and opinion words, was first recognized. Two persons manually categorized the keywords into five product aspects for each review category. The aspect detection performance for each of the SJAOM, sLDA, JST, ASUM, and LARA models was then evaluated against the golden standard using Rand Index on each review domain.

For the evaluations of aspect-specific opinion identification and overall opinion rating prediction, we followed the previous work [PLV02, JO11, LHER12], and considered only two opinion orientations, positive and negative, for an easy and meaningful comparison.

[7] liu.cs.uic.edu/download/data
[8] www.amazon.com
[9] sifaka.cs.uiuc.edu/wang296/Data/index.html
[10] www.tripadvisor.com

In particular, to evaluate the aspect-specific opinion prediction, the two humans then annotated a set of aspect-specific opinion orientations for each test review on a domain. All the five generative models were then evaluated against this ground truth data.

We did not manually create the golden standard data for the overall opinion rating prediction evaluation. Instead, we just used the overall review ratings collected in the data sets, which are in the form of 1 to 5 stars. On Amazon [11], customer reviews with 1-star, 2-star, or 3-star ratings are considered in the critical review category, while reviews with 4-star or 5-star ratings are listed in the favorable review category. Following that, for the given testing data set from each review domain, we converted the lower star ratings of 1, 2, and 3 to negative polarity, and the higher ratings of 4 and 5 to positive polarity. The converted opinion orientations of test reviews from the domain were then used to evaluate the overall opinion prediction for all the six methods, namely, SJAOM, sLDA, JST, ASUM, SVM, and Lexicon.

In the following sections, we will describe each of the three evaluations in detail.

5.4.3 Aspect Detection

We first evaluated aspect detection for SJAOM and baseline models sLDA, JST, ASUM, and LARA on game, CD, and hotel data sets. For each review domain, we showed the best performance in Rand Index given the aspect count of 5 ($K = 5$), as shown in Table 5.3. Note that the baseline methods SVM and Lexicon were not included because they cannot discover the latent topical structure of review data.

SJAOM achieves the best Rand Index score of 83.21% on game review data set, which is 7.77%, 9.56%, 12.86%, 17.18% better than that of baseline models sLDA, JST, ASUM, and LARA, respectively. SJAOM also leads to the highest Rand Index value of 80.19% on CD category, which is 6.98%, 8.1%, 11.28%, and 12.69% better than sLDA, JST,

[11] www.amazon.com

Table 5.3: Rand Index of aspect detection for different models (K=5).

Category	SJAOM	sLDA	JST	ASUM	LARA
Game	83.21%	75.44%	73.65%	70.35%	66.03%
CD	80.19%	73.21%	72.09%	68.91%	67.50%
Hotel	78.07%	72.11%	70.37%	69.29%	63.68%
Average	80.49%	73.59%	72.04%	69.52%	65.74%

ASUM, and LARA. On hotel reviews, SJAOM again results in the best Rand Index of 78.07%, which is 5.96%, 7.7%, 8.78%, and 14.39% better than sLDA, JST, ASUM, and LARA, respectively. We also showed the average aspect detection performance of each model across all three categories in the last line of Table 5.3.

Next, we listed one example aspect detected via SJAOM on each review category, as shown in Table 5.4.

Table 5.4: Example aspects detected via SJAOM on different categories (K=5).

"gameplay" @ Game		"sound" @ CD		"staff" @ Hotel	
game	great	voice	powerful	staff	friendly
gameplay	good	vocals	beautiful	people	helpful
graphics	easy	sound	strong	service	nice
story	fun	songs	emotional	smile	attentive
storyline	enjoyable	talent	haunting	waiters	polite
characters	boring	piano	eerie	management	terrible
music	stupid	range	weird	bartenders	rude
controls	bad	quality	raw	waitress	poor
line	annoying	ballads	spooky	server	flip
plot	hard	melody	distorted	employees	fresh

Each of columns 1, 3 and 5 lists the factual aspect terms covered by an aspect on a review domain, whole each column 2, 4, or 6 shows the semantically related opinion words, where the top 5 words are positive sentiments, and the following 5 words are negative sentiments. Different from benchmark models as well as most other existing topic models, SJAOM does not mix up the clustered aspect term keywords and the related opinion words, instead, it clearly organizes the aspect keywords into different lists of the same cluster. This is due to the special design of separately modeling the aspect term and opinion word in each opinion pair in the SJAOM framework.

We showed an aspect "gameplay" on game data set. The factual keywords like "story", "storyline", "plot", "characters", and "controls", etc., are indicative of the gameplay aspect of a video game, which reflect the way that the game is designed and played, while the opinion words grouped together with keywords such as "easy", "enjoyable", "boring", and "hard", are also semantically related to the "gameplay" aspect. An aspect "sound" was listed on CD category, which covers factual aspect terms like "voice", "vocals", and "sound", as well as the semantically related opinion words like "powerful", "beautiful", and "distorted". We showed the keywords of the aspect "staff" in columns 5 and 6 for hotel review category.

From Tables 5.4, we can see that the factual aspect terms are coherent and meaningful, and are specific enough as well as ratable in online product reviews, while the clustered opinion words are semantically related to the aspect very well.

The proposed supervised joint topic model SJAOM leads to the improved aspect (topic) detection results, compared to the major baseline models for the following reasons:

- The incorporation of overall ratings of reviews into the modeling framework enables SJAOM to exploit the supervision knowledge to guide the inference process of hidden aspects and aspect-based opinion orientations.

- We reduce each review document to a bag of opinion pairs, and feed the pairs into SJAOM. SJAOM can thus model simultaneously each pair of aspect terms and opinion words and to better group semantically related aspect and opinion keywords together. However, all other baselines, sLDA, JST, ASUM, and LARA do not gain from this novel-designed modeling structure for latent aspect/topic detection.

5.4.4 Aspect-specific Opinion Identification

In this section, we tested the aspect-specific opinion identification for SJAOM, sLDA, JST, ASUM, and LARA on the three review categories. We showed the best performance in accuracy for each model given aspect number $K = 5$, as shown in Table 5.5. Note that for sLDA, we used the predicted overall opinion of a review as the identified aspect-based opinion orientations in the review, since no aspect-based opinion analysis layer is designed in sLDA.

Table 5.5: Accuracy of aspect-specific opinion identification on different categories (K=5).

	SJAOM	sLDA	JST	ASUM	LARA
Game	76.24%	59.32%	68.25%	64.54%	63.64%
CD	70.95%	54.41%	63.98%	61.00%	62.96%
Hotel	69.44%	52.38%	63.16%	58.82%	58.15%
Average	72.21%	55.37%	65.13%	61.45%	61.58%

SJAOM results in the best accuracy of 76.24% on game data set, which is 16.92%, 7.99%, 11.7%, and 12.6% better than that of sLDA, JST, ASUM, and LARA, respectively. On CD data set, SJAOM again achieves the highest accuracy of 70.95%, which is 16.54%, 6.97%, 9.95%, and 7.99% better than sLDA, JST, ASUM, and LARA. SJAOM reaches the best accuracy of 69.44% on hotel review domain, which is 17.06%, 6.28%, 10.62%, and 11.29% better than sLDA, JST, ASUM, and LARA, respectively.

The baseline sLDA model results in the worst performance for aspect-specific opinion prediction compared to all other models. This is due to the lack of an opinion analysis layer in its modeling structure. According to experimental results, it is unacceptable to simplistically use the overall opinion of a review as the fine-grained aspect-specific opinion orientations in the review. In practice, it is very common that users comment positively on a product aspect, while expressing negative sentiments toward another aspect of the same product, due to the different qualities of the aspects.

The baselines JST and ASUM are weakly supervised in the sense that a pre-compiled sentiment dictionary is incorporated in their modeling structures to guide the detection of latent opinion orientations. Different from SJAOM, they do not model the supervised overall rating information, and the detection of latent aspects is unsupervised for both of them, thus leading to undesirable performance.

Similar to SJAOM, LARA also incorporates overall review rating data in its structure, but it uses the overall rating information not for identifying latent aspect ratings but for estimating aspect weights. Though achieving comparable performance with ASUM, LARA loses out to our proposed SJAOM model.

On the other hand, SJAOM models overall rating information into the unification framework, and jointly detects underlying aspects and opinion orientations expressed in individual reviews. The model can thus infer which positive or negative aspects contribute to the overall review rating of a product in the reviews. Fortunately, online customer reviews often come with overall ratings, typically in the form of 1 to 5 stars, and it is thus not expensive to collect such supervised data to train the proposed supervised joint topic model SJAOM.

5.4.5 Overall Opinion Prediction

In this section, we evaluated the overall opinion prediction for SJAOM as well as for baseline methods sLDA, JST, ASUM, SVM, and Lexicon. We showed the best accuracy performance (with error bars) versus the number of latent aspects on different review categories. The reported results for the generative models are averaged over five runs, while for each of the non-generative models SVM and Lexicon, the performance does not change with varying the aspect number, thus only one accuracy value is shown.

Figure 5.2 plots the overall opinion prediction results on game data set. Across all eight aspect numbers (from 5 to 40), SJAOM outperforms all benchmark models sLDA,

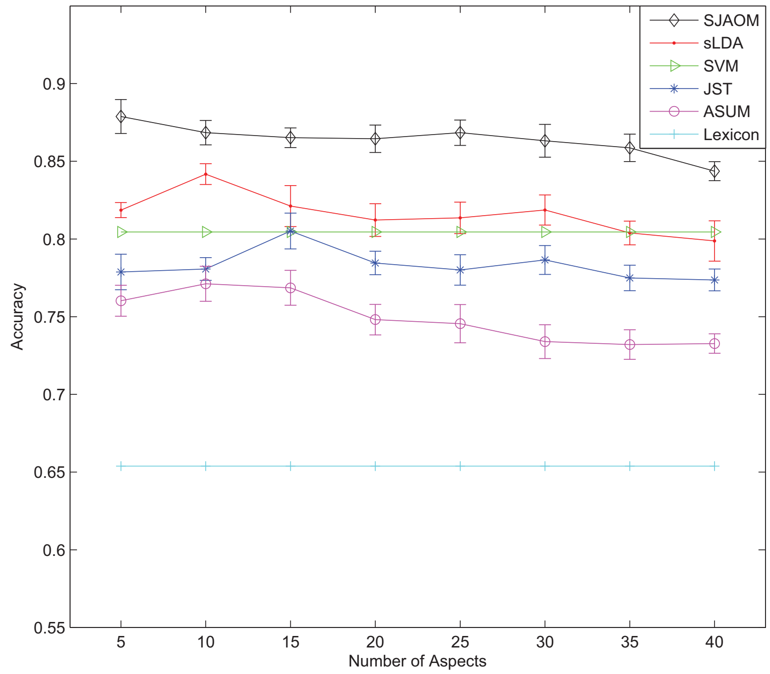

Figure 5.2: Accuracy of overall opinion prediction versus aspect number on game data set.

JST, ASUM, SVM, and Lexicon. SJAOM achieves the best overall opinion prediction accuracy of 87.88% located at aspect number $K = 5$, which is 3.71%, 7.37%, 10.77%, 7.43%, and 22.50% higher than that of sLDA (at aspect number 10), JST (15), and ASUM (10), SVM, and Lexicon, respectively. The SJAOM curve goes down a little with increasing number of aspects.

Figure 5.3 plots the overall opinion prediction performance on CD data set. Again, SJAOM outperforms all the baseline models. As the number of aspect grows to $K = 15$, SJAOM achieves the best accuracy of 82.26%, which is 5.16%, 8.06%, 13.13%, 6.9%, and 24.29% better than that of sLDA (15), JST (5), ASUM (25), SVM, and Lexicon, respectively. After that, the SJAOM curve decreases. At aspect number $K = 35$, it improves a little but was still lower than its best performance.

Figure 5.4 plots the opinion prediction results on hotel data set. The best accuracy of SJAOM is 81.19% at aspect number 10, which is 5.94%, 9.31%, 12.28%, 7.3%, and 22.77%

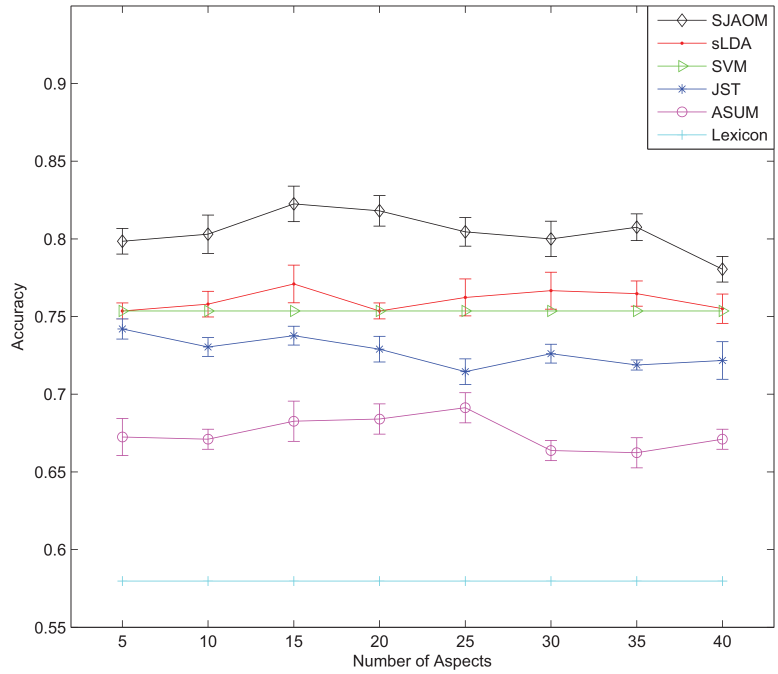

Figure 5.3: Accuracy of overall opinion prediction versus aspect number on CD data set.

better than that of sLDA (25), JST (10), ASUM (20), SVM, and Lexicon, respectively.

According to the experimental results on the data sets, the proposed SJAOM outperforms all the benchmark models, a supervised topic model sLDA, two weakly supervised topic models JST and ASUM, and two more non-generative models SVM and Lexicon.

The unsupervised lexicon-based method (Lexicon) results in the worst overall opinion prediction performance. This is because it tends to suffer from informal expressions frequently mentioned in online customer reviews, such as colloquial slangs and misspelled words, which may not be defined in the sentiment lexicon. Benefitting from the sentiment lexicon, and moreover, the well-designed topic modeling structure, JST and ASUM improve largely the overall opinion prediction performance compared to the Lexicon method. However, the two weakly supervised topic models still lose the contest compared to the fully supervised models.

SVM sentiment classification method results in the improved performance compared to the unsupervised or weakly supervised methods, but performs worse than the super-

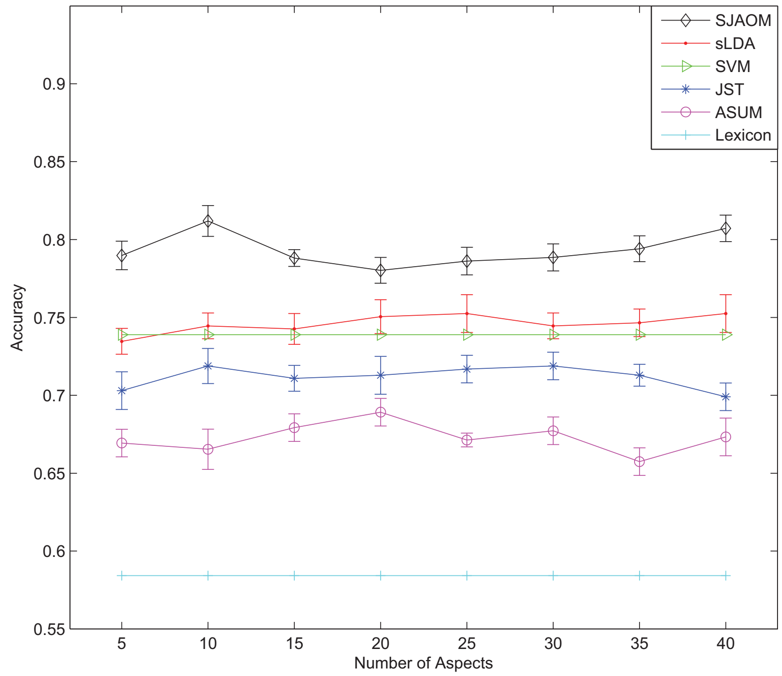

Figure 5.4: Accuracy of overall opinion prediction versus aspect number on hotel data set.

vised topic modeling methods sDLA and SJAOM. One illustration is that supervised topic models benefit from supervised dimensionality reduction, while SVM cannot model the latent topical structure of review data, and thus does not gain from this.

Though sLDA performs well for overall opinion prediction compared to all other baselines, it loses out to our proposed supervised joint topic model SJAOM. The improved effectiveness of SJAOM over sLDA can be attributed to the specially-designed modeling structure of SJAOM for aspect-based review mining and analysis problems:

(i) We exploit the sentiment prior information for better detection of hidden opinion orientations;

(ii) We separate the aspect terms and opinion words in each review, and simultaneously model pairwise aspect terms and opinion words to accurately detect the latent aspects and opinion orientations on the aspects;

(iii) By jointly modeling the factual aspects and the aspect-specific opinion orientations, we can figure out specifically which positive or negative aspects contribute to the overall review ratings.

5.5 Summary

In this chapter, we focus on modeling customer reviews with objectives of identifying semantic aspects of a reviewed entity and opinion orientations on the aspects as well as predicting the overall ratings of individual reviews in a unified framework. We incorporate the observed overall review ratings as supervision data. We also reduce each review document as a bag of opinion pairs to simultaneously model aspect terms and the corresponding opinion words in each review. We thus propose a supervised joint topic model called SJAOM to address the problems. One key advantage of SJAOM is that it can jointly identify the semantic aspects and opinion orientations that are predictive of the overall review ratings.

We conducted experiments using publicly available real-world reviews collected from Amazon and TripAdvisor. We compared SJAOM against six existing representative baselines, a supervised sLDA model, three partially supervised topic models JST, ASUM, and LARA, and two more non-generative models SVM and Lexicon. For the aspect detection and the aspect-specific opinion identification problems, SJAOM results in improved effectiveness against all the generative benchmark models, sLDA, JST, ASUM, and LARA. Note that SVM and Lexicon method cannot model latent topical structure of data, and were not included for the evaluations. As for the overall opinion prediction, SJAOM again outperforms the benchmark methods sLDA, JST, ASUM, SVM, and Lexicon.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this thesis, we focus on the problem of aspect-based opinion mining and analysis of customer reviews. Our objective is to study and develop computational opinion mining techniques to support users to digest the huge volume of raw review data.

In Chapter 1, we described the background and motivation of our research. We also briefly presented three specific problems to be addressed and the methods proposed for them, followed by a summary of the contributions of this thesis.

In Chapter 2, we described the concepts of opinion mining and fine-grained aspect-based opinion mining, and also provided the definitions of terminologies commonly used in the field. We conducted a detailed literature survey on opinion mining as well as aspect-based opinion mining.

In Chapter 3, we studied the problem of extracting aspect terms and opinion words that appear in reviews. We proposed a generalized corpus statistics association based bootstrapping framework (ABOOT) for the problem. Two particular statistical association measures, likelihood ratio tests (LRT) and latent semantic analysis (LSA), are evaluated in the framework, leading to two instance methods of ABOOT, LRTBOOT and LSABOOT. Experimental results using real-world reviews demonstrate the benefits of the proposed LRTBOOT and LSABOOT over the benchmark methods. In fact, LRTBOOT can achieve surprisingly good performance with only one seed, which is simply the domain word, e.g., “cellphone” for cellphone reviews. This makes our proposed association-based bootstrapping approach powerful and effective for aspect term and opinion word extraction in the practical opinion analysis of online reviews.

In Chapter 4, we studied the problem of identifying implicit aspects for the opinion words devoid of explicit aspects in reviews. We proposed a two-phrase cooccurrence association rule mining method (coARM) for the problem. To the best of our knowledge, this is the first work that conducts quantitative evaluation for implicit aspect identification. The experimental results on real-life reviews demonstrate that corARM outperforms the well-established baseline methods.

In Chapter 5, we dealt with modeling customer reviews with the objectives of identifying the semantic aspects of a reviewed entity and opinions on the aspects as well as predicting the overall ratings in a unified framework. We proposed a novel supervised joint topic model called SJAOM for the problem. One key advantage of SJAOM is its ability to identify the hidden aspects and opinions that are predictive of the overall ratings of reviews. Experimental results show the improved effectiveness of SJAOM over six existing representative benchmark models, a supervised sLDA model, three partially supervised topic models JST, ASUM, and LARA, and two more non-generative models SVM and Lexicon.

6.2 Future Directions

For future work, we present several promising directions of our research on opinion mining.

6.2.1 Review Quality Evaluation

Fine-grained aspect-based opinion analysis and summarization results have been useful for users to make informed decisions. In reality, people may also want to read the most helpful original reviews to learn the first-hand experiences and opinions from others. It is thus necessary and important to develop a review quality evaluation system to automatically select the most helpful reviews to users.

Some e-commerce websites already provide a crowd-source mechanism to evaluate review quality. For instance, Amazon allows customers to vote each product review as *helpful* or *unhelpful*. However, a good many of reviews receive very few or no votes at all. Decisions made using the sparse voting information tend to suffer from bias.

By formulating the review utility evaluation as a regression or classification problem, previous studies focus on developing different sets of data features, and then train a review utility evaluation function on the features. Zhang and Varadarajan [ZV06] developed regression model on textual features to predict the helpfulness of online reviews. Liu et al. [LHAY08] developed regression model using a different set of text features for the review utility evaluation. Ghose and Ipeirotis [GI11] proposed to exploit text-related features, such as subjectivity, readability, and linguistic correctness, to estimate the helpfulness of product reviews. Further, to improve the textual feature based review quality predictor, Lu et al. [LTNP10] proposed to exploit social contextual information about reviewers' identities and social networks.

All existing works focus on predicting the overall helpfulness/utility of a review for the reviewed entity (entity-level), and they do not address the problem of predicting fine-grained utilities of a review for individual aspects of the entity. Thus, they fail to select the most helpful reviews for fine-grained entity aspects.

For future work, we plan to deal with review helpfulness/utility prediction not only at product level but also at aspect level. Then, based on the estimated review utility scores, a list of the most helpful reviews can be selected for a reviewed entity as well as for each aspect of the entity. A new supervised generative probabilistic model will be developed for the problem. It is able to jointly model the product-level and aspect-level review utilities, and forms the prediction of the review utilities via regression process in the unified framework.

6.2.2 Opinion Spam Detection

Generally, existing studies on opinion mining and sentiment analysis assume that their opinion resources are genuine and truthful. As a matter of fact, people nowadays game the review system via generating fake reviews and opinions to promote or demote products or services. Opinion spam detection simply refers to detecting fake reviews, review spammers, and spammer groups.

Faked reviews [JL08], also known as deceptive opinions [OCCH11], are unworthy positive or malicious negative reviews that have been deliberately created to deceive people. The unworthy positive reviews are typically generated to promote products or services, while the malicious negative reviews are generated primarily to defame the reputations of competing products or services. Review spammer [LNJ+10], or spammer in short, indicates the reviewer who generate faked reviews. Spammer group [MLG12] refers to a group of spammers who works together writing fake reviews.

One key challenge of opinion spam detection lies in that it is very hard to recognize fake reviews by manually reading them, which makes it difficult to collect ground truth spam data to develop opinion spam detection systems [Liu12].

Jindal and Liu [JL08] detected duplicate reviews from a given review corpus. They used the duplicates as fake reviews and the rest of the reviews as non-fake reviews in the training data. Then, they developed review-centric, reviewer-centric, and product-centric features, and built a logistic regression model to discriminate the fake reviews from non-fake ones. Based on insights coming from research in psychology and computational linguistics, Ott et al. [OCCH11] developed three approaches to deceptive/fake review detection. They found that while standard n-gram based text categorization was the best individual detection approach, a combination approach using psycholinguistically motivated features and n-gram features can perform slightly better.

Lim et al. [LNJ+10] addressed the problem of detecting review spammers, or finding reviewers who are the sources of fake reviews. They developed four spamming behavior models/indicators relying on different spamming patterns of review content and ratings. Each model/indicator assigns a numeric spamming behavior score to a reviewer by measuring the extent to which the reviewer practises spamming behavior of a certain type. They finally assigned an overall spam score to each review by combining the spam scores of different spamming behavior models.

In reality, it is common that a group of reviewers work together to promote or to demote products or services. Mukherjee et al. [MLG12] studied the problem of spammer group detection. They first used a frequent itemset mining method to generate a list of candidate spammer groups. They then defined group spam behavior indicators as well as individual spam behavior indicators. Based on the indicators, they developed group spam-products model, member spam-product model, and group spam-member spam model to capture the inter-relationship among products, groups, and group members. They finally introduced a group spam ranking algorithm to iteratively measure the spamicity of each candidate spam group by employing the three relation models.

For future work, we plan to detect fake reviews. We will develop a semi-supervised leaning method for the problem. One key advantage of this method is that the model training can exploit a large number of unlabelled reviews in addition to a small set of annotated review data. Moreover, a stochastic optimization method will be also developed for dealing with large scale review data. Furthermore, we also plan to develop an unsupervised generative probabilistic model to detect hidden fake reviewer groups and to compute the spamicity of each reviewer group via a regression procedure based on the membership and spamicity of each reviewer in the group.

6.2.3 Opinion Analysis via Modeling Multiple Resources

Online reviews often come with location and time information. It will be much meaningful to mine the geographical and temporal characteristics of opinions and sentiments in practice. This is because users could make more wiser decisions based on the informative geographical and temporal opinion analysis results. We plan to extend our proposed supervised joint topic model by jointly modeling the location and temporal data to provide dynamic and location-aware opinion analysis on a entity (product) or specific aspects of the entity.

Some e-commerce websites, such as CNET [1], Epinions [2], and TripAdvisor [3], also provide additional useful information, such as pre-fixed review format, e.g., Pros and Cons, a pre-specified list of entity aspects, and aspect-specific ratings, etc. Such prior knowledge can be exploited to improve the aspect-based opinion analysis performance. We will also consider modeling such prior information in the extended model.

[1] www.cnet.com
[2] www.epinions.com
[3] www.tripadvisor.com

Appendix A

Derivation of Gibbs Sampling Equation for SJAOM

In this appendix, we derive the Gibbs sampling Equation 5.3 for the supervised joint aspect and opinion model (SJAOM) proposed in Chapter 5.

First, we list the distributions used in the derivation as following:

- $\psi_k \sim \text{Dir}(\lambda)$: Dirichlet distribution over aspect words
- $\phi_{kl} \sim \text{Dir}(\beta)$: Dirichlet distribution over opinion words
- $\theta_m \sim \text{Dir}(\alpha)$: Dirichlet distribution over aspects
- $\pi_{mk} \sim \text{Dir}(\gamma)$: Dirichlet distribution over opinion orientations
- $a_{mn} \sim \text{Mult}(\theta_m)$: Multinomial distribution over aspect assignment
- $o_{mn} \sim \text{Mult}(\pi_{ma_{mn}})$: Multinomial distribution over opinion assignment
- $t_{mn} \sim \text{Mult}(\psi_{a_{mn}})$: Multinomial distribution over aspect words
- $w_{mn} \sim \text{Mult}(\phi_{a_{mn}o_{mn}})$: Multinomial distribution over opinion words
- $r_m \sim \text{N}(\eta^T \bar{z}_m, \sigma^2)$: Normal linear model over overall rating response, where

$$\bar{z}_m = \frac{1}{C}\sum_{n=1}^{N_m}(a_{mn} \times (\omega^T \times o_{mn})).$$

For an opinion pair $\langle t_{mn}, w_{mn} \rangle$ with the index $i = (m, n)$, we are interested in deriving the following full conditional distribution:

$$p(a_i = k, o_i = l \mid \boldsymbol{a}^{\neg i}, \boldsymbol{o}^{\neg i}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2),$$

where, the quantities $\boldsymbol{a}^{\neg i}$ and $\boldsymbol{o}^{\neg i}$ indicate the assignments of aspects and opinion orientations to all aspect terms and opinion words in the corpus except for the assignments a_i and o_i for the aspect term and opinion word at position i, respectively.

We expand the conditional distribution as follows:

$$\begin{aligned} & p(a_i = k, o_i = l \mid \boldsymbol{a}^{\neg i}, \boldsymbol{o}^{\neg i}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2) \\ & = \frac{p(\boldsymbol{a}, \boldsymbol{o}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2)}{p(\boldsymbol{a}^{\neg i}, \boldsymbol{o}^{\neg i}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2)} \\ & \propto p(\boldsymbol{a}, \boldsymbol{o}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2). \end{aligned} \tag{A.1}$$

In SJAOM, the joint distribution of aspect terms, opinion words, the assignments of aspects and opinion orientations, and overall ratings can be factored as follows:

$$\begin{aligned} & p(\boldsymbol{a}, \boldsymbol{o}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2) \\ & = p(\boldsymbol{a} \mid \alpha) \cdot p(\boldsymbol{o} \mid \boldsymbol{a}, \gamma) \cdot p(\boldsymbol{t} \mid \boldsymbol{a}, \lambda) \cdot p(\boldsymbol{w} \mid \boldsymbol{a}, \boldsymbol{o}, \beta) \cdot p(\boldsymbol{r} \mid \boldsymbol{a}, \boldsymbol{o}, \eta, \sigma^2). \end{aligned} \tag{A.2}$$

For the first term in Equation A.2, integrating out θ, we yield:

$$\begin{aligned} p(\boldsymbol{a} \mid \alpha) &= \int p(\boldsymbol{a} \mid \theta) p(\theta \mid \alpha) d\theta \\ &= \int \prod_m \prod_k \theta_{m,k}^{N_{m,k}} \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_{m,k}^{\alpha_k - 1} d\theta_m \\ &= \prod_m \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int \prod_k \theta_{m,k}^{N_{m,k} + \alpha_k - 1} d\theta_m \\ &= \prod_m \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(N_{m,k} + \alpha_k)}{\Gamma(\sum_k (N_{m,k} + \alpha_k))} \\ &= \prod_m \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(N_{m,k} + \alpha_k)}{\Gamma(N_m + \sum_k \alpha_k)}, \end{aligned} \tag{A.3}$$

where $N_{m,k}$ indicates the count of words in document (review) m assigned to aspect k, N_m is the total number of words in document m, and $\Gamma(x)$ means a Gamma function.

Integrating out π for the second term in Equation A.2, we obtain:

$$
\begin{aligned}
p(\boldsymbol{o} \mid \boldsymbol{a}, \gamma) &= \int p(\boldsymbol{o} \mid \boldsymbol{a}, \pi) p(\pi \mid \gamma) d\pi \\
&= \int \prod_m \prod_k \prod_l \pi_{m,k,l}^{N_{m,k,l}} \frac{\Gamma(L\gamma)}{\Gamma(\gamma)^L} \prod_l \pi_{m,k,l}^{\gamma-1} d\pi \\
&= \prod_m \prod_k \frac{\Gamma(L\gamma)}{\Gamma(\gamma)^L} \int \prod_l \pi_{m,k,l}^{N_{m,k,l}+\gamma-1} d\pi \\
&= \prod_m \prod_k \frac{\Gamma(L\gamma)}{\Gamma(\gamma)^L} \frac{\prod_l \Gamma(N_{m,k,l}+\gamma)}{\Gamma(\sum_l (N_{m,k,l}+\gamma))} \\
&= \prod_m \prod_k \frac{\Gamma(L\gamma)}{\Gamma(\gamma)^L} \frac{\prod_l \Gamma(N_{m,k,l}+\gamma)}{\Gamma(N_{m,k}+L\gamma))},
\end{aligned}
\tag{A.4}
$$

where $N_{m,k,l}$ is the count of words in document m assigned to aspect k and opinion l.

We integrate out ψ for the third term in Equation A.2 and obtain:

$$
\begin{aligned}
p(\boldsymbol{t} \mid \boldsymbol{a}, \lambda) &= \int p(\boldsymbol{t} \mid \boldsymbol{a}, \psi) p(\psi \mid \lambda) d\psi \\
&= \int \prod_k \prod_u \psi_{k,u}^{N_{k,u}} \frac{\Gamma(U\lambda)}{\Gamma(\lambda)^U} \prod_u \psi_{k,u}^{\lambda-1} d\psi \\
&= \prod_k \frac{\Gamma(U\lambda)}{\Gamma(\lambda)^U} \int \prod_u \psi_{k,u}^{N_{k,u}+\lambda-1} d\psi \\
&= \prod_k \frac{\Gamma(U\lambda)}{\Gamma(\lambda)^U} \frac{\prod_u \Gamma(N_{k,u}+\lambda)}{\Gamma(\sum_u (N_{k,u}+\lambda))} \\
&= \prod_k \frac{\Gamma(U\lambda)}{\Gamma(\lambda)^U} \frac{\prod_u \Gamma(N_{k,u}+\lambda)}{\Gamma(N_k+U\lambda))},
\end{aligned}
\tag{A.5}
$$

where $N_{k,u}$ is the number of unique aspect word u in vocabulary assigned to aspect k across all documents, and N_k indicate the sum of $N_{k,u}$ across all aspect words in vocabulary.

Integrating out ϕ for the fourth term in Equation A.2, we yield:

$$
\begin{aligned}
p(\boldsymbol{w} \mid \boldsymbol{a}, \boldsymbol{o}, \beta) &= \int p(\boldsymbol{w} \mid \boldsymbol{a}, \boldsymbol{o}, \phi) p(\phi \mid \beta) d\phi \\
&= \int \prod_k \prod_l \prod_v \phi_{k,l,v}^{N_{k,l,v}} \frac{\Gamma(\sum_v \beta_{l,v})}{\prod_v \Gamma(\beta_{l,v})} \prod_v \phi_{k,l,v}^{\beta_{l,v}-1} d\phi \\
&= \prod_k \prod_l \frac{\Gamma(\sum_v \beta_{l,v})}{\prod_v \Gamma(\beta_{l,v})} \int \prod_v \phi_{k,l,v}^{N_{k,l,v}+\beta_{l,v}-1} d\phi \\
&= \prod_k \prod_l \frac{\Gamma(\sum_v \beta_{l,v})}{\prod_v \Gamma(\beta_{l,v})} \frac{\prod_v \Gamma(N_{k,l,v} + \beta_{l,v})}{\Gamma(\sum_v (N_{k,l,v} + \beta_{l,v}))} \\
&= \prod_k \prod_l \frac{\Gamma(\sum_v \beta_{l,v})}{\prod_v \Gamma(\beta_{l,v})} \frac{\prod_v \Gamma(N_{k,l,v} + \beta_{l,v})}{\Gamma(N_{k,l} + \sum_v \beta_{l,v})},
\end{aligned}
\tag{A.6}
$$

where $N_{k,l,v}$ is the count of unique opinion word v in vocabulary assigned to opinion l and aspect k, and $N_{k,l}$ is the sum of $N_{k,l,v}$ across all the opinion words in vocabulary.

We expand the last term in Equation A.2 as follows:

$$
\begin{aligned}
p(\boldsymbol{r} \mid \boldsymbol{a}, \boldsymbol{o}, \eta, \sigma^2) &= \prod_m p(r_m \mid a_m, o_m, \eta, \sigma^2) \\
&= \prod_m p(r_m \mid \bar{z}_m, \eta, \sigma^2) \\
&= \prod_m \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(r_m - \eta^T \bar{z}_m)^2}{2\sigma^2}).
\end{aligned}
\tag{A.7}
$$

Next, we substitute Equations A.3, A.4, A.5, A.6, and A.7 into Equation A.2. We then cancel out the terms of Equation A.2 which are independent of the aspect term t_{mn} and opinion word w_{mn} as well as the aspect and opinion assignments a_{mn} and o_{mn}. We thus yield the full conditional distribution, on which the Gibbs sampling algorithm draws the latent aspect a_i and opinion orientation o_i for an opinion pair $\langle t_{mn}, w_{mn} \rangle$ with the index i:

$$
\begin{aligned}
& p(a_i = k, o_i = l \mid \boldsymbol{a}^{\neg i}, \boldsymbol{o}^{\neg i}, \boldsymbol{t}, \boldsymbol{w}, \boldsymbol{r}, \alpha, \gamma, \lambda, \beta, \eta, \sigma^2) \\
& \propto \frac{N_{m,k}^{\neg i} + \alpha_k}{N_m^{\neg i} + \sum_{k\prime} \alpha_{k\prime}} \cdot \frac{N_{m,k,l}^{\neg i} + \gamma}{N_{m,k}^{\neg i} + L\gamma} \cdot \frac{N_{k,u}^{\neg i} + \lambda}{N_k^{\neg i} + U\lambda} \\
& \cdot \frac{N_{k,l,v}^{\neg i} + \beta_{l,v}}{N_{k,l}^{\neg i} + \sum_{v\prime} \beta_{l,v\prime}} \cdot \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(r_m - \eta^T \bar{z}_m^{\neg i})^2}{2\sigma^2}).
\end{aligned}
\tag{A.8}
$$

References

[ABM09] Apoorv Agarwal, Fadi Biadsy, and Kathleen R. Mckeown. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL'09, pages 24–32, Athens, Greece, 2009.

[AIS93] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, SIGMOD'93, pages 207–216, Washington, D.C., United States, 1993.

[BE10] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT'10, pages 804–812, Los Angeles, California, 2010.

[BHV04] Philip Beineke, Trevor Hastie, and Shivakumar Vaithyanathan. The sentimental factor: Improving review classification via human-provided information. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL'04, Barcelona, Spain, 2004.

[BM07] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Proceedings of the 21st Annual Conference on Neural Information Processing Systems - Volume 7*, pages 121–128, Vancouver, Canada, 2007.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[BPP96] Adam L Berger, Vincent J Della Pietra, and Stephen A Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[CLL10] Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: a chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING'10, pages 13–16, Beijing, China, 2010.

[CVXS06] Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs*, pages 27–29, 2006.

[DDH90] Scott Deerwester, Susan T. Dumais, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society For Information Science*, 41:391–407, 1990.

[DLP03] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW'03, pages 519–528, Budapest, Hungary, 2003.

[DLY08] Xiaowen Ding, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the International Conference on Web Search and Data Mining*, WSDM'08, pages 231–240, Palo Alto, California, USA, 2008.

[Dun93] Ted Dunning. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, 19(1):61–74, March 1993.

[ES05] Andrea Esuli and Fabrizio Sebastiani. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM'05, pages 617–624, Bremen, Germany, 2005.

[GEM09] Gayatree Ganu, Noemie Elhadad, and Amelie Marian. Beyond the stars: Improving rating predictions using review text content. In *Proceedings of the Twelfth International Workshop on the Web and Databases*, 2009.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, 1984.

[GI11] A. Ghose and P.G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, 2011.

[GS04a] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[GS04b] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Science*, volume 101, pages 5228–5235, Jan 2004.

[GVL96] G.H. Golub and C.F. Van Loan. Matrix computations. *The Johns Hopkins University Press*, 3rd edition, 1996.

[GZ06] Andrew B. Goldberg and Xiaojin Zhu. Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 45–52, 2006.

[GZG+09] Honglei Guo, Huijia Zhu, Zhili Guo, XiaoXun Zhang, and Zhong Su. Product feature categorization with multilevel latent semantic association. In *Proceeding of the 18th ACM conference on Information and knowledge management*, CIKM'09, pages 1087–1096, Hong Kong, China, 2009.

[HCSK10] Zhen Hai, Kuiyu Chang, Qinbao Song, and Jung-jae Kim. A statistical nlp approach for feature and sentiment identification from chinese reviews.

In *Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China*, pages 105–112, 2010.

[He10] Yulan He. Learning sentiment classification model from labeled features. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM'10, pages 1685–1688, Toronto, ON, Canada, 2010.

[HKH+01] Vasileios Hatzivassiloglou, Judith L. Klavans, Melissa L. Holcombe, Regina Barzilay, Min yen Kan, and Kathleen R. McKeown. Simfinder: A flexible clustering tool for summarization. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 41–49, 2001.

[HL04a] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD'04, pages 168–177, Seattle, WA, USA, 2004.

[HL04b] Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pages 755–760, San Jose, California, 2004.

[HLA11] Yulan He, Chenghua Lin, and Harith Alani. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT'11, pages 123–131, Portland, Oregon, 2011.

[HM97] Vasileios Hatzivassiloglou and Kathleen R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL'98, pages 174–181, Madrid, Spain, 1997.

[Hof99] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, UAI'99, pages 289–296, Stockholm, Sweden, 1999.

[HW00] Vasileios Hatzivassiloglou and Janyce M. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics - Volume 1*, COLING'00, pages 299–305, Germany, 2000.

[JG10] Niklas Jakob and Iryna Gurevych. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'10, pages 1035–1045, Cambridge, Massachusetts, 2010.

[JHS09] Wei Jin, Hung Hay Ho, and Rohini K. Srihari. Opinionminer: A novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'09, pages 1195–1204, Paris, France, 2009.

[JL08] Nitin Jindal and Bing Liu. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Data Mining*, WSDM'08, pages 219–230, Palo Alto, California, USA, 2008.

[JO11] Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM'11, pages 815–824, Hong Kong, China, 2011.

[Joa98] Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, pages 137–142, 1998.

[KC07] Lun-Wei Ku and Hsin-Hsi Chen. Mining opinions from the web: Beyond relevance retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 58(12):1838–1850, October 2007.

[KH04] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING'04, Geneva, Switzerland, 2004.

[KH06] Soo-Min Kim and Eduard Hovy. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST'06, pages 1–8, Sydney, Australia, 2006.

[KHC09] Lun-Wei Ku, Ting-Hao Huang, and Hsin-Hsi Chen. Using morphological and syntactic structures for chinese opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1260–1269, 2009.

[KI06] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2):110–125, 2006.

[KLWC05] Lun-Wei Ku, Li-Ying Lee, Tung-Ho Wu, and Hsin-Hsi Chen. Major topic detection and its application to opinion summarization. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 627–628, 2005.

[KM03] Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL'03, pages 423–430, Sapporo, Japan, 2003.

[KT02] Jon Kleinberg and Eva Tardos. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *J. ACM*, 49(5):616–639, September 2002.

[Lew98] David D. Lewis. Naive bayes at forty: The independence assumption in information retrieval. In *Proceedings of the 10th European Conference on Machine Learning*, ECML'98, pages 4–15, 1998.

[LH09] Chenghua Lin and Yulan He. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM'09, pages 375–384, Hong Kong, China, 2009.

[LHAY08] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Modeling and predicting the helpfulness of online reviews. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*, pages 443–452, 2008.

[LHC05] Bing Liu, Minqing Hu, and Junsheng Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th International Conference on World Wide Web*, WWW'05, pages 342–351, Chiba, Japan, 2005.

[LHER12] Chenghua Lin, Yulan He, Richard Everson, and Stefan Ruger. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145, June 2012.

[LHH+10] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. Structure-aware review mining and summarization. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 653–661, Beijing, China, 2010.

[LHZL10] Shoushan Li, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL'10, pages 414–423, Uppsala, Sweden, 2010.

[Liu10] Bing Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2010.

[Liu12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, May 2012.

[LLJ+11] Fangtao Li, Nathan Nan Liu, Hongwei Jin, Kai Zhao, Qiang Yang, and Xiaoyan Zhu. Incorporating reviewer and product information for review rating prediction. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, pages 1820–1825, 2011.

[LMP01] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML'01, pages 282–289, 2001.

[LNJ+10] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM'10, pages 939–948, Toronto, ON, Canada, 2010.

[LTNP10] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, WWW'10, pages 691–700, Raleigh, North Carolina, USA, 2010.

[LZS09] Tao Li, Yi Zhang, and Vikas Sindhwani. A non-negative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL'09, pages 244–252, Suntec, Singapore, 2009.

[MBF+90] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244, 1990.

[MC04] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of the conference on*

Empirical methods in natural language processing, volume 4 of *EMNLP'04*, pages 412–418, 2004.

[ME11] Samaneh Moghaddam and Martin Ester. Ilda: Interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'11, pages 665–674, Beijing, China, 2011.

[MGL09] Prem Melville, Wojciech Gryc, and Richard D. Lawrence. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'09, pages 1275–1284, Paris, France, 2009.

[Min00] Thomas Minka. Estimating a dirichlet distribution. In *Technical report, MIT*, 2000.

[MLG12] Arjun Mukherjee, Bing Liu, and Natalie Glance. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st International Conference on World Wide Web*, WWW'12, pages 191–200, Lyon, France, 2012.

[NDA06] Vincent Ng, Sajib Dasgupta, and S. M. Niaz Arifin. Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL'06, pages 611–618, Sydney, Australia, 2006.

[OCCH11] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT'11, pages 309–319, Portland, Oregon, 2011.

[PE05] Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT'05, pages 339–346, Vancouver, British Columbia, Canada, 2005.

[PL04] Bo Pang and Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL'04, Barcelona, Spain, 2004.

[PL05] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL'05, pages 115–124, Ann Arbor, Michigan, 2005.

[PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP'02, pages 79–86, 2002.

[PT12] Georgios Paltoglou and Mike Thelwall. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Trans. Intell. Syst. Technol.*, 3(4):66:1–66:19, September 2012.

[QIW10] Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 913–921, Beijing, China, 2010.

[QLBC11] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Comput. Linguist.*, 37:9–27, 2011.

[QWB+08] Guang Qiu, Can Wang, Jiajun Bu, Kangmiao Liu, and Chun Chen. Incorporate the syntactic knowledge in opinion mining in user-generated content. In *WWW2008 Workshop: NLP Challenges in the Information Explosion Era*, Beijing, China, 2008.

[Ran71] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[Ril96] Ellen Riloff. Automatically generating extraction patterns from untagged text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, AAAI'96, pages 1044–1049, Portland, Oregon, 1996.

[RPW06] Ellen Riloff, Siddharth Patwardhan, and Janyce Wiebe. Feature subsumption for opinion analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'06, pages 440–448, Sydney, Australia, 2006.

[RW03] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'03, pages 105–112, 2003.

[RWW03] Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL'03, pages 25–32, Edmonton, Canada, 2003.

[SC08] Veselin Stoyanov and Claire Cardie. Topic identification for fine-grained opinion analysis. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING'08, pages 817–824, Manchester, United Kingdom, 2008.

[SM08] Vikas Sindhwani and Prem Melville. Document-word co-regularization for semi-supervised sentiment analysis. In *Proceedings of the Eighth IEEE International Conference on Data Mining*, ICDM'08, pages 1025–1030, 2008.

[SXG+08] Qi Su, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. Hidden sentiment association in chinese web opinion mining. In *Proceedings of the 17th international conference on World Wide Web*, WWW'08, pages 959–968, Beijing, China, 2008.

[TBT+11] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307, June 2011.

[Tes59] Lucien Tesniere. *Elements de Syntaxe Structurale*. Librairie C. Klincksieck, Paris, 1959.

[TL03] Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Trans. Inf. Syst.*, 21(4):315–346, October 2003.

[TM08a] Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceeding of the 17th international conference on World Wide Web*, WWW'08, pages 111–120, Beijing, China, 2008.

[TM08b] Ivan Titov and Ryan T. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 308–316, Columbus, Ohio, 2008.

[TNK10] Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823–848, 2010.

[Tur02] Peter D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, pages 417–424, Philadelphia, Pennsylvania, 2002.

[TWC08] Songbo Tan, Yuefen Wang, and Xueqi Cheng. Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'08, pages 743–744, Singapore, Singapore, 2008.

[WGA05] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, CIKM'05, pages 625–631, Bremen, Germany, 2005.

[Wie00] Janyce Wiebe. Learning subjective adjectives from corpora. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, pages 735–740, 2000.

[WLZ10] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'10, pages 783–792, Washington, DC, USA, 2010.

[WLZ11] Hongning Wang, Yue Lu, and ChengXiang Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'11, pages 618–626, San Diego, California, USA, 2011.

[WR05] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the Sixth International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497, 2005.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language*

Processing, HLT'05, pages 347–354, Vancouver, British Columbia, Canada, 2005.

[WWH09] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.

[WZHW09] Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. Phrase dependency parsing for opinion mining. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Volume 3 - Volume 3*, EMNLP'09, pages 1533–1541, Singapore, 2009.

[XLL+13] Liheng Xu, Kang Liu, Siwei Lai, Yubo Chen, and Jun Zhao. Mining opinion words and opinion targets in a two-stage framework. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1764–1773, Sofia, Bulgaria, 2013.

[YC11] Ainur Yessenalina and Claire Cardie. Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, pages 172–182, Edinburgh, United Kingdom, 2011.

[YH03] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the conference on Empirical methods in natural language processing*, EMNLP'03, pages 129–136, 2003.

[YNBN03] Jeonghee Yi, Tetsuya Nasukawa, Razvan Bunescu, and Wayne Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Proceedings of the Third IEEE International Conference on Data Mining*, ICDM'03, pages 427–434, 2003.

[YYC10] Ainur Yessenalina, Yisong Yue, and Claire Cardie. Multi-level structured models for document-level sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'10, pages 1046–1056, Cambridge, Massachusetts, 2010.

[ZC08] Taras Zagibalov and John Carroll. Automatic seed word selection for unsupervised sentiment classification of chinese text. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING'08, pages 1073–1080, Manchester, United Kingdom, 2008.

[ZLLOS10] Lei Zhang, Bing Liu, Suk Hwan Lim, and Eamonn O'Brien-Strain. Extracting and ranking product features in opinion documents. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING'10, pages 1462–1470, Beijing, China, 2010.

[ZLW08] Jun Zhao, Kang Liu, and Gen Wang. Adding redundant features for crfs-based sentence sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'08, pages 117–126, Honolulu, Hawaii, 2008.

[ZLXJ10] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. Grouping product features using semi-supervised learning with soft-constraints. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING'10, pages 1272–1280, Beijing, China, 2010.

[ZLXJ11] Zhongwu Zhai, Bing Liu, Hua Xu, and Peifa Jia. Clustering product features for opinion mining. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM'11, pages 347–354, Hong Kong, China, 2011.

[ZV06] Zhu Zhang and Balaji Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM'06, pages 51–57, Arlington, Virginia, USA, 2006.