אוניברסיטת בן-גוריון בנגב
הפקולטה למדעי הטבע
המחלקה למדעי המחשב

שימוש בכותרות LDA בזיהוי וניתוח של נושאים וסנטימנט בתוך הטקסטים

חיבור זה מהווה חלק מהדרישות לקבלת התואר "מוסמך למדעי טבע" (M.Sc.)

מאת: מאשה איגרא
בהנחיית: פרופ' מיכאל אלחדד

טבת תשע"ג ינואר 2013

אוניברסיטת בן-גוריון בנגב
הפקולטה למדעי הטבע
המחלקה למדעי המחשב

שימוש בכותרות LDA בזיהוי וניתוח של נושאים וסנטימנט בתוך הטקסטים

חיבור זה מהווה חלק מהדרישות לקבלת התואר "מוסמך למדעי טבע" (M.Sc.)

מאת: מאשה איגרא
מנחה: פרופ' מיכאל אלחדד

חתימת הסטודנט: _____ תאריך:_____
חתימת המנחה: _____ תאריך: _____
חתימת יו"ר הועדה המחלקתית: _____ תאריך:_____

טבת תשע"ג ינואר 2013

תקציר בעברית:

עם גידול דרמטי של מאגרי תוכן באינטרנט הגיע צורך לכלים אוטומטיים המסוגלים להפיק מידע רלוונטי למשתמש מטקסט רגיל. אחד הנושאים הנחקרים ביותר בעיבוד טקסט אוטומטי הוא ניתוח סנטימנט של הטקסט: רגש שמביע כותב מסוים כלפי נושא כלשהוא. סנטימנט יכול להיות חיובי, שלילי או ניטרלי.
אנו חוקרים את שיטות האוטומטיות של ניתוח הסנטימנט של הטקסט וכיצד מבנה התוכן יכול להשפיע ולשפר את ניתוח הסנטימנט. אנו נותנים סקירה של טכניקות שונות המנתחות את הסנטימנט של הנושאים המתוארים בטקסט. אנחנו מציעים שיטה דו-שלבית חדשה לזיהוי וניתוח סנטימנט של נושאי הטקסט. בעקבות אינטואיציה לשונית ובעקבות מחקרים קודמים מידע הקשרי עשיר יכול לתרום לניתוח של טקסטים כגון מציאת נושאים בתוך הטקסט וסיווג רגשות.
השיטה שלנו מתבצעת בשני שלבים עצמאיים, זיהוי נושאי הטקסט וסיווג ההיבט הרגש. בשני השלבים, אנו משתמשים בשיטת סיווג מבוקרת SVM על מנת לסווג טקסטים לפי נושאים ולפי סנטימנט ובודקים האם שימוש בנושאים המוסקים מן הטקסט בצורה לא מבוקרת (LDA) יכול לשפר גילוי של סנטימנט של הטקסט.
בשלב ראשון אנחנו מחלצים את נושאי הטקסט ובשלב השני, עבור כל אחד מנושאים, אנחנו מסווגים כל אחד מהטקסטים לפי הסנטימנט.
על מנת לבנות מסווגים טובים יותר עבור סיווג הרגש, אנחנו מציעים להשתמש בתכוניות המוסקות בצורה לא מבוקרת (LDA) ביחד עם תכוניות של מילים בטקסט.
תכוניות המוסקות בצורה לא מבוקרת שאנחנו משתמשים בהן הן משקלות של כל אחד מהנושאים המוסקים מתוך כל משפט בטקסט. על מנת להסיק את התכוניות הללו, אנחנו משתמשים בשיטה לוקלית של מודל הסתברותי גנרטיבי הנקרא LDA.

BEN- GURION UNIVERSITY OF THE NEGEV
THE FACULTY OF NATURAL SCIENCES
DEPARTMENT OF COMPUTER SCIENCES

Use of LDA Topics in Aspect and Sentiment Analysis

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE MASTER OF SCIENCES DEGREE

Masha Igra
UNDER THE SUPERVISION OF: Prof. Michael Elhadad

January 2013

BEN- GURION UNIVERSITY OF THE NEGEV
THE FACULTY OF NATURAL SCIENCES
DEPARTMENT OF COMPUTER SCIENCES

Use of LDA Topics in Aspect and Sentiment Analysis

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE MASTER OF SCIENCES DEGREE

Masha Igra
UNDER THE SUPERVISION OF: Prof. Michael Elhadad

Signature of student: _____ Date: _____
Signature of supervisor: _____ Date: _____
Signature of chairperson
Of the committee for graduate studies: _____ Date:_____

January 2013

Abstract

With the dramatic growth of user generated content comes a corresponding need for automatic tools capable of extracting relevant information for the user from plain text. We investigate sentiment analysis and how content structure can benefit sentiment text analysis. We give an overview of various techniques used to tackle the problems in the domain of sentiment analysis. We propose a two-step approach for aspect sentiment classification. Following linguistic intuition and previous research, we demonstrate that rich contextual information can benefit text analysis and improve accuracy on aspect extraction and sentiment classification. LDA-based topic models provide an operational set of features that contribute to the accuracy of each of the two stages: aspect identification and sentiment analysis on a large range of datasets. Our approach proceeds in two independent steps, aspect identification and sentiment classification. For both steps, we use a supervised SVM-based classification approach and we evaluate whether topic models (in the form of LDA models inferred from the data) provide effective guidance to each classifier. The *first* step attempts to extract the aspects of an object. The *second* step classifies sentiment over each of these aspects.

In order to construct better classifiers for sentiment classification, we propose to use unsupervised features along with bag-of-words features. Our unsupervised features are topics distribution over inferred aspects for each sentence in the data. In order to infer the salient aspects in the data, we specifically use topics distributions inferred by a local version of probabilistic generative model (LDA).

Table of contents

Chapter 1:Introduction

Sentiment Analysis is a part of Natural Language Processing (NLP) that deals with the automated extraction of opinions of an author of written texts. The goal is to decide automatically whether an author has a positive or negative sentiment/opinion towards a certain concept or aspect of a concept.

In this context, we refer to the following definition of "opinion":

> "An opinion is simply a positive or negative sentiment, view, attitude, emotion, or appraisal about an entity or an aspect of the entity from an opinion holder." [Kim and Hovy, 2004]

Sentiment analysis is the task of classifying a part of a document (review) into positive or negative classes. One might think this is an easy task, and hypothesize that the polarity of opinions can generally be identified by a set of keywords. But the results of early studies on movie reviews classification suggest that coming up with the right set of keywords might be less trivial than one might initially think. Consider the following example:

> "This film should be *brilliant*. It sounds like a *great* plot, the actors are *first grade*, and the supporting cast is *good* as well, and Stallone is attempting to deliver a *good* performance. However, it can't hold up."

As indicated by the emphasis, words that are positive in orientation dominate this example, and yet the overall sentiment is negative because of the negative last sentence. Sentiment can often be expressed in a more subtle manner, no ostensibly negative words occur, e.g., "She runs the gamut of emotions from A to B."

Online review sites continue to grow in popularity as more people seek the advice of fellow users regarding services and products. Unfortunately, users are often forced to wade through large quantities of written data in order to find the information they want. This has led to an increase in research in the areas of opinion mining and sentiment analysis, with the aim of providing systems that can automatically analyze user reviews and extract the information most relevant to the user.

One example of such an application is generating a summary of the important factors mentioned in the reviews of a product. Another application is comparing two similar products. In this case, it is important to present to the user the aspects in which the products differ, rather than just provide a general star rating. A third example is systems for generating automatic recommendations, based on similarity between products, user reviews, and history of previous purchases. These types of applications require an underlying framework to identify the important aspects of the product (also known as features or attributes), and the sentiment expressed by the review writer.

From the perspective of a user reading the reviews to get information about a product, the evaluations of the specific aspects are just as important as the overall rating of the product.

A user looking to buy a digital camera may want to know what a review says about the photo quality, brightness of lens, and shutter speed of a Panasonic, not just whether the review recommends the camera. Although sometimes the aspect information is available, it is unlikely to be a comprehensive set of all aspects that are evaluated in the reviews.

Another important task in review analysis is discovering how opinions and sentiments for different aspects are expressed. The cell phone's battery lasts “long", a laptop's screen “reflects”, and a restaurant's server is “attentive". Overall review sentiment doesn’t express a sentiment for each product aspect and sometimes can be opposite to a specific aspect sentiment.

In this work, we review automatic methods for sentiment analysis and aspect identification. The task we consider takes as input raw text segmented in sentences. Each sentence is classified in terms of aspects (what is the topic of the sentence) and sentiment (what sentiment is expressed relative to the aspect). Several standard datasets have been developed in the past few years to benchmark solutions to this type of task. We review these datasets and the types of solutions proposed in the past. We then present our method: we proceed in two independent steps, aspect identification and sentiment classification. For both steps, we use a supervised SVM-based classification approach and we evaluate whether topic models (in the form of LDA models inferred from the data) provide effective guidance to each classifier. Our experiments demonstrate that this approach achieves high accuracy over a wide range of product review datasets.

Chapter 2:Objectives

This work studies the following general research question: can there be automatic recognition of sentiment from text?

A basic task in sentiment analysis is determining the overall polarity of the opinions of a given text in the document. Sentiment opinion can be expressed by a user as positive, negative, or neutral or with a ratable score.

Mostly, opinion documents express opinion/sentiment on different products/aspects, where the aspect (or feature) is an attribute associated with a product (entity). Opinions expressed on different aspects can hold different polarities and in such a case, an overall sentiment cannot represent true sentiment. Most readers prefer to get an aspect sentiment score in order to compare it to other products.

Our goal is feature/aspect-based sentiment analysis. This refers to determining the opinions or sentiments expressed on different features or aspects of entities, e.g., of a cell phone, a digital camera, or a bank. A feature or aspect is an attribute or component of an entity, e.g., the screen of a cell phone or the picture quality of a camera. For example: "*I bought an iPhone a few days ago. It is such a nice phone, although a little large. The touch screen is cool. The voice quality is clear too. I simply love it!*" In this example, the product is iPhone, and its ratable aspects are *screen*, *voice*, and *size*, and an overall score is provided in addition to a sentiment on each aspect.

Sentiment analysis research works on unstructured texts of global social media, such as reviews, forum discussions, blogs, and social networks. All of these text forms include public opinions. Our research focuses on review datasets. Each review can contain several opinions. Some of them are positive opinions, while others are negative or emotions. Reviews can refer to specific aspects. Each review can be expressed by one opinion holder or different opinion holders. In general, opinions can be expressed about anything, e.g., a product, a service, an individual, an organization, an event, or a topic, by any person or organization.

Our research focused on review datasets in the English language. Each review contains from 1 to 100 sentences, with an average length of 10 sentences. Our semi-supervised approach requires labeled datasets: aspects and aspect-sentiment tags per each aspect-sentiment review. The objective is for a given collection of opinionated documents to output all discovered aspects and their sentiments, and compare the results to known baselines and other approaches.

In this context, we are addressing the following questions: How accurately can we reveal different aspects from unstructured text? Which sentiment analysis approach is more effective: 2-phase approach or joint approach? Which knowledge sources are more effective: words, syntax relations, words, or aspects?

Chapter 3:Datasets and Evaluation Metrics

We review in this Chapter a set of publically available datasets in the field of sentiment analysis that have been used in previous research. These datasets indicate the type of phenomena we are interested in identifying.

3.1. Datasets Description

3.1.1. Restaurants Dataset [Ganu et al., 2009]

The corpus contains 5,531 restaurants, with associated structured information (location, cuisine type) and a set of reviews. There are 52,264 reviews, of which 1,359 are editorial reviews, the rest are user reviews. Reviews contain structured metadata (star rating, date) along with text. Typically reviews are small; the reviews are written by different users. The data set is sparse: restaurants typically have only a few reviews, with 1388 restaurants having more than 10 reviews; and users typically review few restaurants, with only 299 (non-editorial) users having reviewed more than 10 restaurants.

The following six categories were identified: Food, Service, Price, Ambience, Anecdotes, and Miscellaneous. The first four categories are typical parameters of restaurant reviews. Anecdotal sentences are sentences describing the reviewer's personal experience or context, but that do not usually provide information on the restaurant quality (e.g., "*I knew upon visiting NYC that I wanted to try an original deli.*"). The Miscellaneous category captures sentences that do not belong to the other five categories and includes sentences that are general recommendations (e.g., "*Your friends will thank you for introducing them to this gem!*"). One sentence can contain more than one aspect category. In addition to sentence categories, sentences have an associated sentiment: Positive, Negative, Neutral, or Conflict. Users often seem to compare and contrast good and bad aspects; this mixed sentiment is captured by the Conflict category (e.g., "*The food here is rather good, but only if you like to wait for it.*").

Table 1: Sentence distribution over aspects.

Aspect	**Number of sentences**
Anecdote	8,922
Food	28,692
Price	5,783
Miscellaneous	20,758
Ambience	9,203
Staff	14,096
Total	80,000

Table 2: Sentence distribution over sentiments.

Sentiment	**Number of sentences**
Positive	129,652
Neutral	26,642
Negative	37,182
Conflict	7,804

3.1.2. Hotels Dataset [Baccianella et al., 2009]

The dataset is a set of 15,763 hotel reviews obtained by crawling from the TripAdvisor Web site. Each review has a score of one to five "stars", both globally and for each of seven aspects: "BusinessService" (Bservice), "CheckIn/FrontDesk", "Cleanliness", "Location", "Rooms", "Service", and "Value". Aside from the "global" dataset, seven aspect-specific datasets are defined, which contain the reviews for which a label has been attributed for the given aspect (not all reviews contain scores for all of the aspects).

Table 3: Distribution of sentences per aspect.

	1-score	**2-score**	**3-score**	**4-score**	**5-score**	**Total**
Service	361	524	1,175	2,590	4,288	8,938
Bservice	199	244	925	869	874	3,111
Check in	202	247	728	1,234	2,839	5,250
Clean	113	319	815	2,502	5,276	9,025
Location	88	261	594	1,535	2,763	5,241
Rooms	313	599	1,450	3,518	3,132	9,012
Value	401	560	1,568	3,209	3,156	8,894
Total	1,677	2,754	7,255	15,457	22,328	49,471

3.1.3. Multi-Domain Sentiment Dataset [Blitzer et al., 2007]

The Multi-Domain Sentiment Dataset is constructed by selecting Amazon product reviews for four different product types: books, DVDs, electronics, and kitchen appliances. Each review consists of a rating (0–5 stars), a reviewer name and location, a product name, and the review text. Reviews with rating > 3 were labeled positive, those with rating < 3 were labeled negative, and the rest discarded because their polarity was ambiguous. This dataset is balanced between positive and negative examples.

Table 4: Review distribution over domains.

Domain	**No. of Reviews**
Books	840
DVD	305
Electronics	1,350
Kitchen	1,189
Total	3,684

3.1.4. DVD Reviews [Sauper et al., 2010]

DVD reviews were obtained from the website IGN.com. Each review is accompanied by 1–10 scale ratings in four categories that assess the quality of a movie's content, video, audio, and DVD extras. In this data set, segments corresponding to each of the aspects are clearly delineated in each document. Each review in the data file is listed as a set of aspects, e.g.:

{"extras": (score, sentences), "audio": (score, sentences),
"video": (score, sentences), "movie": (score, sentences)}.

Therefore, we can compare the performance of the algorithm using automatically induced content models against the gold standard structural information.

Table 5: Distribution over aspects.

Aspect	No. of Reviews
Movie	665
Audio	665
Video	665
Extras	665
Total	2660

3.2. Evaluation Metrics

The datasets described above were evaluated using the following metrics:

- **Accuracy** – the proportion of the total number of predictions that were correct: AC = (TP + TN)/(all retrieved reviews)
- **Recall** – the proportion of positive cases that were correctly identified: R = TP/(TP + FN)
- **Precision** – the proportion of the predicted positive cases that were correct: P = TP/(TP + FP)
- **F1** is a measure of a test's accuracy:

  F1 = $2 * \frac{precision * recall}{precision + recall}$
- **F2** is a measure of a test's accuracy that weights recall higher than precision:

  F2 = $(1 + 2^2) * \frac{precision * recall}{2^2 * precision + recall}$
- **Precision at n (P@n)** with n equal to 5, 10, and 20. Precision at n is a percentage of relevant words among first N top words retrieved.
- **Mean absolute difference (L1)** between system prediction (SP) and true 1–10 sentiment ratings (SR) across test documents and aspects. *n* is a number of measures.

  L1 = $\frac{\sum | SP - SR |}{n}$
- **Mean squared difference (L2)** between system prediction and true 1–10 sentiment ratings across test documents and aspects.

  L2 = $\frac{\sum | SP - SR |^2}{n}$
- **Kendall's tau coefficient (τk) of opinion words** [Jijkoun and Hofmann, 2009]**.** This measure looks at the number of pairs of ranked items that agree or disagree with the ordering in the gold standard. The value of τk ranges from -1 (perfect disagreement) to 1 (perfect agreement), with 0 indicating an almost random ranking.
- **Kendall's distance (Dk) of opinion words** [Jijkoun and Hofmann, 2009] is similar to Kendall's tau coefficient; the value ranges from 0 (perfect agreement) to 1 (perfect disagreement).

3.3. Existing Results

We describe which metrics were used for each dataset in order to evaluate aspect detection and sentiment analysis.

3.3.1. Aspects Detection Metrics

Table 6: Aspects detection metrics per data set

Data set name	Accuracy	Recall	Precision	F1	P@n	L1	L2	τk	Dk
Restaurants		[Zhao et al., 2010]	[Zhao et al., 2010]	[Zhao et al., 2010]					
Hotels		[Brody and Elhadad, 2010]	[Brody and Elhadad, 2010]						
Multi-Domain									
DVD									

3.3.2. Sentiment Analysis Metrics

Table 7: Sentiment analysis metrics per data set

Data set name	Accuracy	Recall	Precision	F1	P@n	L1	L2	τk	Dk
Restaurants					[Zhao et al., 2010]	[Baccianella et al., 2009]		[Brody & Elhadad, 2010]	[Brody & Elhadad, 2010]
Hotels					[Zhao et al., 2010]				
Multi-Domain									
DVD						[Sauper et al., 2010]	[Sauper et al., 2010]		

Because every dataset is evaluated with different metrics, it is difficult to compare their results.

Chapter 4:Previous Work

4.1. Knowledge Sources for Sentiment Analysis

Most existing techniques for document-level sentiment classification are based on supervised learning, although there are also some unsupervised methods. In most sentiment analysis approaches, the following features have been used:

- *Terms and their frequency*: These features are individual words or word n-grams and their frequency counts. In some cases, word positions may also be considered. The TF-IDF weighting scheme from information retrieval may be applied too. These features have been shown to be quite effective in sentiment classification.
- *Part of speech*: It was found in much research that adjectives are important indicators of opinions. Thus, adjectives have been treated as special features.
- *Opinion words and phrases*: Opinion words are words that are commonly used to express positive or negative sentiments. For example, *beautiful*, *good*, and *amazing* are positive opinion words, and *bad*, *poor*, and *terrible* are negative opinion words. Apart from individual words, there are also opinion phrases and idioms, e.g., "*cost someone an arm and a leg*". In order to collect the opinion word list, three main approaches have been investigated: manual, dictionary-based and corpus-based approach. The manual approach is very time-consuming and thus it is not usually used alone, but combined with automatic approaches.
    - Dictionary-based approach: One of the simple techniques in this approach is based on bootstrapping using a small set of seed opinion words and an online dictionary, e.g., Wordnet [Miller et al., 1990] or thesaurus [Mohammad et al., 2006]. The strategy is to first collect a small set of opinion words manually with known orientation, and then to grow this set by searching in the WordNet or thesaurus for their synonyms and antonyms.
    - Corpus-based approach: The methods in the corpus-based approach rely on syntactic or co-occurrence patterns and also a seed list of opinion words to find other opinion words in a large corpus. One of the key ideas technique starts with a list if seed opinion adjectives, and uses them and a set of linguistic constraints or conventions on connections to identify additional adjectives opinion words and their orientation.
- *Negations*: Clearly, negation words are important because their appearances often change the opinion orientation. For example, the sentence *"I don't like this camera"* is negative.
- *Syntactic dependency*: Word dependency-based features generated from parsing or dependency trees have also been tried by several researchers.

4.2. Two-phase Approach

There are two major approaches of aspect-grained sentiment analysis:

1) Two-phase approach: The first phase attempts to extract the aspects of an object that users frequently rate. The second phase classifies and aggregates sentiment over each of these aspects.

2) Joint model of a content model (aspects discovery) and a sentiment analysis. The joint model discovers aspects and sentiment simultaneously.

4.2.1. Aspect Detection

The earliest attempts at aspect detection were based on the classic information extraction approach of using frequently occurring noun phrases (e.g., [Hu and Liu 2004]). Such approaches work well in detecting aspects that are strongly associated with a single noun, but are less useful when aspects encompass many low frequency terms (e.g., the *food* aspect of restaurants, which involves many different dishes), or are abstract (e.g., *ambiance* can be described without using any concrete nouns at all).
Wang et al present a bootstrapping-based algorithm to identify the major aspects (guided by a few seed words describing the aspects) [Wang et al., 2010]. Given the seed words for each aspect and all the review text as input, each sentence is assigned to the aspect that shares the maximum term overlapping with this sentence; words with high dependencies into the corresponding aspect are included in the keyword list. These steps are repeated until the aspect keyword list is unchanged.

Common solutions to the task of aspect detection involve clustering with the help of knowledge-rich methods, involving manually constructed rules, semantic hierarchies, or both (e.g., [Popescu et al., 2005]).

The standard topic modeling approach – Latent Dirichlet Allocation (LDA) [Blei et al., 2003] – is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document. LDA uses the bag-of-words representation of documents; therefore it explores co-occurrences at the document level and discovers overall topics of a document. LDA has been used for aspect detection in several manners. All the variants described below attempt to identify fine-grained ratable topics from the text, using different variants of the basic LDA algorithm. Topic modeling is a popular tool for aspect identification because it is intuitively appealing and works in an unsupervised manner over a noisy document set. The details of operation, however, can be tricky, and many different models have been proposed in the past few years. We briefly describe 6 such variants in the following.

One of the first proposed models for building topics that are representative of ratable aspects is the Multi-Grain topic model (MG-LDA) [Titov and McDonald, 2008], which attempts to capture two layers of topics: global and local, where the local topics correspond to ratable aspects. For example, consider an extract from a review of a London hotel: "… *public transport in London is straightforward, the tube station is about an 8 minute walk ... or you can get a bus for £1.50".* It can be viewed as a mixture of topic *London* shared by the entire review (words: "*London", "tube", "£"*) and the ratable aspect location, specific for the local context of the sentence (words: *"transport", "walk", "bus*"). The proposed model includes two distinct types of topics: global topics and local topics. The hypothesis is that ratable aspects will be captured by local topics and global topics will capture properties of reviewed items. They represent a document as a set of sliding windows, each covering *T* adjacent sentences within it. Each window *v* in document *d* has an associated distribution over local topics $\theta_{d,v}^{loc}$ and a distribution defining preference for local topics versus global topics $\pi_{d,v}$. A word can be sampled using any window covering its sentence *s*, where the window is chosen according to a categorical distribution ψ_s .

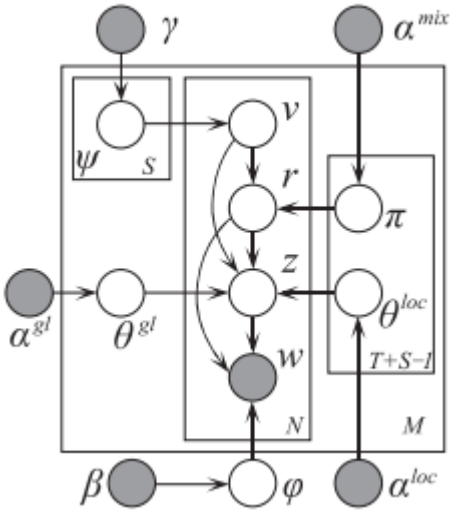

Figure 1: MG-LDA model

The formal definition of the model with K^{gl} global and K^{loc} local topics is the following. First, draw K^{gl} word distributions for global topics φ_z^{gl} from a Dirichlet prior $Dir(\beta^{gl})$ and K^{loc} word distributions for local topics $\varphi_{z'}^{loc}$ from $Dir(\beta^{loc})$. Then, for each document *d*:

- Choose a distribution of global topics $\theta_d^{gl} \sim Dir(\alpha^{gl})$.
- For each sentence *s* choose a distribution $\psi_{d,s}(v) \sim Dir(\gamma)$.
- For each sliding window *v*
  - Choose $\theta_{d,v}^{loc} \sim Dir(\alpha^{loc})$,
  - Choose $\pi_{d,v} \sim Beta(\alpha^{mix})$.
- For each word *i* in sentence *s* of document *d*
  - Choose window $v_{d,i} \sim \psi_{d,s}$,
  - Choose $r_{d,i} \sim \pi_{d,v_{d,i}}$
  - If $r_{d,i} = gl$ choose global topic $z_{d,i} \sim \theta_d^{gl}$,
  - If $r_{d,i} = loc$ choose local topic $z_{d,i} \sim \theta_{d,v_{d,i}}^{loc}$,
  - Choose word $w_{d,i}$ from the word distribution $\varphi_{z_{d,i}}^{r_{d,i}}$.

As an evaluation dataset Titov and McDonald used a subset of reviews for Mp3 players from Google Product Search and subsets of reviews of hotels and restaurants from Google Local Search. All system runs are evaluated using ranking loss, which measures the average distance between the true and predicted numerical ratings. Overall ranking loss is simply the average over each aspect. A lower loss means a better performance. Their method improved N-gram base line results. But in this model it is hard to map between dozens of extracted local topics and ratable aspects. An extension of this model helps to solve this issue but needs to provide an aspect-specific rating along with the review text.

Brody and Elhadad [Brody and Elhadad, 2010] introduced a method which operates on sentences, rather than documents, and employs a small number of topics that correspond directly to aspects.

Word level aspect extraction [Maas et al., 2011] learned word vectors via an unsupervised probabilistic model of documents (LDA-like). It models word probabilities conditioned on the topic mixture variable. A probabilistic model of a document uses a continuous mixture distribution over words indexed by a multi-dimensional random variable θ. The model assumes each word *w* is conditionally independent of the other words, given θ. Each word is represented as a vector; one could view the entries of a word vector as that word's association strength with respect to each latent topic dimension.

Zhang et al. [Zhang et al., 2011] propose a new probabilistic generative model for topic analysis of online reviews, called Author-Experience-Object-Topic Model (AEOT). This model captures the relationship between the authors, objects, and reviews in order to improve the performance of topic analysis. They modified LDA model by adding an author layer and an object layer. Each word in a review can be generated from an author, the object, or the experience of authors on the object. They compared their discovered topics with LDA topics and found that their topics are more coherent.

Zhai et al. [Zhai et al., 2011] proposed a semi-supervised LDA method, called *constrained-LDA* for topics modeling. They incorporate two types of constraints into the popular topic modeling method LDA: must-link and cannot-link constraints. A must-link constraint specifies that two data instances must be in the same cluster. A cannot-link constraint specifies that two data instances cannot be in the same cluster. All constraints are extracted automatically with no human involvement:

- **Must-link**: If two product features share one or more words, they assume they form a must-link, i.e., they should be in the same topic, e.g., "battery power" and "battery life".
- **Cannot-link**: If two product features occur in the same sentence and they are not connected by "and", the two features form a cannot-link.

The main idea of the proposed approach is to revise the topic updating probabilities computed by LDA using the probabilities induced from the constraints. That is, in the topic updating process, they compute an additional probability q from the must-links and cannot-links for every candidate topic, and then multiply it by the probability calculated by the original LDA model as the final probability for topic updating. Their method performs much better than the original LDA.

Zhan and Li [Zhan et al., 2011] proposed a semantic-dependent word pairs generative model (SDWP) for pairs of nouns and adjectives for each sentence. They extend the LDA model by, instead of dealing just with unigram representation, dealing with pairs of nouns and adjectives with semantic relevance, from which clusters of nouns and adjectives should be obtained for each aspect. They model sentences as pairs of nouns and adjectives with semantic dependence. These semantic dependent pairs are extracted from preprocessed dependency trees. A generative process of SDWP for each sentence samples topics and samples a noun and its semantic neighbor from multinomial probability conditioned on the topic. They compared their approach with standard LDA and reported better results.

4.2.2. Sentiment Analysis

Existing methods of automated sentiment analysis can be classified into three kinds: word-based, sentence-based, and overall document.

Word-based Sentiment Classification

Hatzivassiloglou and McKeown [Hatzivassiloglou and McKeown, 1997] present an approach based on linguistic heuristics. Their technique aims at extracting a list of adjectives that have positive and negative meanings. It relies on the fact that in the case of polarity classification, the two classes of interest represent opposites, and we can utilize "opposition constraints" to help make labeling decisions. Specifically, constraints between pairs of adjectives are induced from a large corpus by observing whether the two words are linked by conjunctions such as "but" and "and". The task is then cast as a clustering or binary-partitioning problem where the inferred constraints are to be obeyed. Once the clustering has been completed, the label "*positive orientation*" is assigned to the class of words whose members have the highest average frequency.

Sentence-based Sentiment Classification

Wilson et al. [Wilson et al., 2005] used word-dependency trees in order to build features for sentence-wise sentiment polarity classification. They used a two-step process: the first step classifies each phrase as neutral or polar; the second step takes polar phrases and analyzes their polarity.

Shaikh et al. [Shaikh et al., 2005], with similar accuracy, performed semantic dependency analysis on the semantic verb frames of each sentence, and applied a set of rules to each dependency relation to calculate the contextual valence of the whole sentence. Their approach relies on the semantic relationship between the structure of natural language and contextual valence of the words used on a given text.

Erikson [Eriksson, 2005] used parsed trees in order to extract features from sentences for a machine learning-based document sentiment classifier. He applied an objective sentence removal algorithm and worked just with subjective sentences. The accuracy of his method was a little bit lower but very close to Shaikh's.

Green and Resnik [Green and Resnik, 2009] learned the connection between the structure of a sentence and implicit sentiment. They suggested that the relationship is mediated by a set of "grammatically relevant" semantic properties well known to be important cross-linguistically in characterizing the interface between syntax and lexical semantics. They used observable proxies for underlying semantic features describing the semantic properties of sentences. In the experiments on death penalty domain, their classifier showed better performance than past classifiers.

Document-based Sentiment Classification

In the latest studies on document sentiment analysis, classifiers based on machine learning, which have been successful in other document classification tasks, showed higher performance than rule-based classifiers.

Pang et al. [Pang et al., 2002] classified movie reviews. They applied a supervised machine learning method to document sentiment classification. They used word N-grams in the dataset as features for their classifier. A word N-gram is a set of N continuous

words. The best results were obtained using a word unigram-based model run through SVMs, reaching 82.9% accuracy.

Pang and Lee [Pang and Lee, 2004] also attempted to improve their method by focusing only on subjective sentences in the reviews. But the accuracy of their method is less than that of the classifier using full reviews.

Dave et al. [Dave et al., 2003] used machine learning methods to classify reviews on several kinds of products. Unlike Pang's research, they reached the best accuracy rate with a word bigram classifier on their dataset. This discrepancy seems to indicate that features are dependent on the data.

Matsumoto et al. [Matsumoto et al., 2005] used word order and syntactic relations between words in a sentence for a machine learning-based document sentiment classifier. They obtained sub-pattern features as information of word order and syntactic relations between words in a document by mining frequent sub-patterns from word sequence and dependency trees in the dataset.
For example:

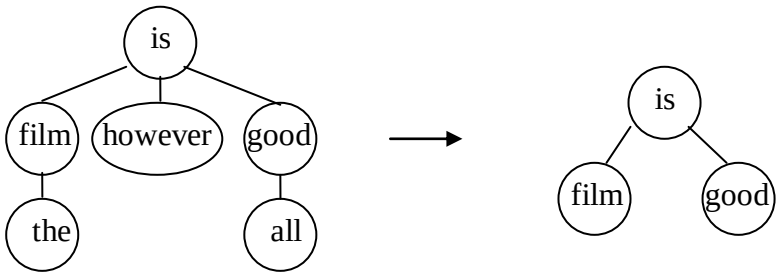

Figure 2: Dependency tree and frequent sub-tree of a sentence: "*the film, however, is all good*".

In the experiments on the movie review domain, they report about 92% accuracy.

Brody and Elhadad [Brody and Elhadad, 2010] detected sentiment classification of adjectives in an unsupervised manner. They developed the following procedure. For each aspect, they extracted the relevant adjectives, built a conjunction graph, automatically determined the seed set (or used a manual one, for comparison), and propagated the polarity scores to the rest of the adjectives. (This method extends the approach introduced in [Hatzivassiloglou and McKeown, 1997] to the context of a per-aspect / per-sentence classification.) The adjective graph introduces a document level structure over sentences in the dataset.

Shivashankar and Ravindran [Shivashankar and Ravindran, 2010] introduced an iterative classification algorithm that performs multi-grain classification in a semi-supervised environment. Intra-dependencies at sentence level are captured using a domain knowledge base, i.e., relation between features of a domain. But this method assumes background domain knowledge that has the set of features and similarity between the features and the lexicon of sentiment terms (say, for instance, General Inquirer). They propose the following procedure:

Procedure: A document is seen as a function of sentence-level labels, and a sentence can in turn be seen as a function of tuple level labels. The entire corpus can be posed as an undirected graph (V, E), where E is the set of edges and the nodes in V correspond to different entities – document, sentence, and tuples. A tuple has target features and sentiment terms, and with the domain knowledge base a neighborhood structure is formed within a document and between documents. Since not all the nodes are labeled, the goal is to utilize the information available in the overall graph structure, and fully label the nodes. The intuition behind the method (Collective classification) is that the node's probability to be assigned a label increases, given that its neighbors are assigned the same label.
Datasets: CNET, Epinions, and Edmunds, which contain Automobile reviews.
Results: They showed results better than their base line (Lexical classifier: classify documents as positive or negative depending on the count of sentiment terms).

Wang and Lu introduced a regression model called Latent Rating Regression (LRR) [Wang et al., 2010], which aims at inferring aspect ratings and weights for each individual review based only on the review content and the associated overall rating. The idea of reviewer's rating behavior is as follows: to generate an opinionated review, the reviewer first decides the aspects she wants to comment on; then, for each aspect, the reviewer carefully chooses the words to express her opinions. The reviewer then forms a rating on each aspect based on the sentiments of words she used to discuss that aspect. Finally, the reviewer assigns an overall rating depending on a weighted sum of all the aspect ratings, where the weights reflect the relative emphasis she has placed on each aspect.

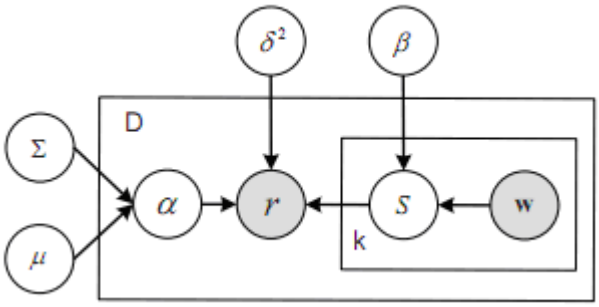

Figure 3: Graphical Representation of LRR. The outer box represents reviews, while the inner box represents the composition of latent aspect rating and word descriptions within a review.

W_d is a word frequency matrix that gives a normalized frequency of words in each aspect. s_d and α_d are review-level k-dimensional aspect weight vector and aspect rating vector, respectively. An aspect rating s for each aspect is generated as a linear combination of W_{di} and βi. r is an overall rating of the review treated as the response variable. In order to capture the dependencies among different aspects, they employ a multivariate Gaussian distribution as the prior for aspect weights, i.e., $\alpha_d \sim N(\mu, \Sigma)$, where μ and Σ are the mean and variance parameters.
This method showed better results in some evaluations metrics compared to base line.

Duric and Song [Duric and Song, 2011] proposed a new feature selection scheme for sentiment analysis that uses an unsupervised learned Context and Syntax model. They train an HMM-LDA model to give them the syntactic classes and semantic (topical) class. For each syntactic class they select the representative words by cumulative probability. For sentiment analysis they just use words related to syntactic classes as most the representative sentiment features in the classifier.

4.3. Joint Models

Joint models detect polarity and topics simultaneously. Intuitively, polarities are dependent on topics or domains. For instance, though the adjective *unpredictable* in a phrase such as "*unpredictable steering*" may have negative orientation in an automobile review, it could also have positive orientation in a phrase like "*unpredictable plot*" in a movie review. Joint models classify polarity and topics on multiple levels: words, sentences, and entire documents.

One of the first joint models based on LDA was the Joint Sentiment/Topic (JST) model [Lin and He, 2009]. Lin and He model document sentiments by adding to the standard LDA an additional sentiment layer between the document and the topic layer. Hence, JST is effectively a four-layer model, where sentiment labels are associated with documents, under which topics are associated with sentiment labels and words are associated with both sentiment labels and topics.

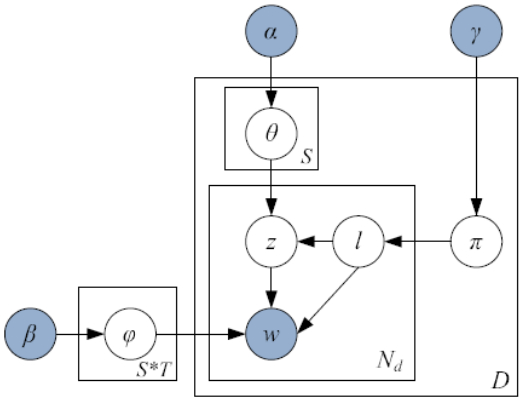

Figure 4: JST model

The formal definition of the generative process that corresponds to the hierarchical Bayesian model shown in Figure 1 is as follows:

- For each document *d*, choose a distribution $\pi_d \sim Dir(\gamma)$.
- For each sentiment label *l* under document *d*, choose a distribution $\theta_{d,l} \sim Dir(\alpha)$.
- For each word *wi* in document *d*
  - Choose a sentiment label $l_i \sim \pi_d$,
  - Choose a topic $z_i \sim \theta_{d,l_i}$,
  - Choose a word *wi* from the distribution of words defined by the topic *zi* and sentiment label *li*, φ_{zi}^{li}.

They used a movie review data set and reported results that were worse than SVM-based approaches, although appealing because the approach is fully unsupervised.

Jo and Oh [Jo and Oh, 2011] introduced a Sentence-LDA (SLDA) and Aspect and Sentiment Unification Model (ASUM) and reported better document-sentiment accuracy results than the previously mentioned JST. Sentence aspect and sentiment classification assumes that one sentence tends to represent one aspect and one sentiment. The following example evaluates several aspects including *price, size,* and *sound*, and each sentence expresses a sentiment about one aspect.

> *"I've owned my computer for almost a week now, and I'm absolutely loving it. For the money I almost bought an Hp with a T6600 processor and 320GB hard drive etc., etc. I bought this for the same price as the other hp and got MORE memory and a better processor.*
>
> *I also love its 14" **monitor**, so I can take it anywhere and fit it into any **bag**. It has a very sleek and glossy exterior. A little bit of work to keep clean, but that's no reason not to buy a computer. The speakers are great for a laptop. I was actually surprised how clear sounds and music are."*

In the first sentence of the second paragraph, the words "*monitor*" and "*bag*" co-occur. In general, these two words are not closely related, but the co-occurrence of them signals that this sentence is evaluating the size of the monitor. This joint classification model can be applied to various tasks:

- Aspect discovery: finds aspects that match the details of the reviews.
- Senti-aspect discovery: finds senti-aspect words that reflect both aspect and sentiment.
- Aspect-specific sentiment words.
- Sentiment classification: In order to determine the sentiment of a review, the probabilistic sentiment distribution in a review can be used, such that a review is set to be positive if positive sentiment has an equal or higher probability than negative sentiment, and set to be negative otherwise.

Sentence-LDA (SLDA) is a probabilistic generative model that assumes all words in a single sentence are generated from one aspect. In **SLDA**, the generative process is as follows:

1. For every aspect *z*, draw a word distribution.
2. For each review *d*,
    (a) Draw the review's aspect distribution;
    (b) For each sentence,
        i. Choose an aspect *z*
        ii. Generate words *w.*

The difference with basic LDA is that the distribution of topics is per sentence (and not per document) and that all words within each sentence are sampled from the same aspect. SLDA is further extended into the Aspect and Sentiment Unification Model (**ASUM**), which incorporates aspect and sentiment to model sentiments toward the different aspects as follows:

- A reviewer first decides to write a review of a restaurant that expresses a distribution of sentiments.
- He decides the distribution of the aspects for each sentiment.
- He decides, for each sentence, a sentiment to express and an aspect for which he feels that sentiment.

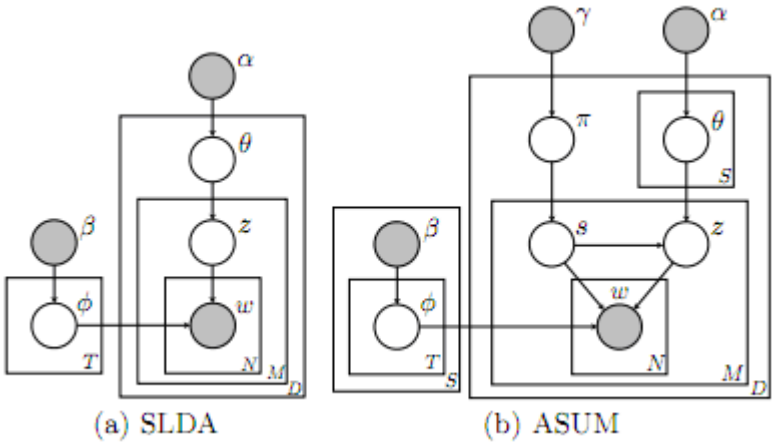

Figure 5: Graphical representation of SLDA and ASUM.

They used electronic device reviews from Amazon and restaurant reviews from the Yelp datasets. The results improved over the JST with reported accuracy on sentiment / aspect classification increased from 0.68 to 0.84.

One of the popular aspects of specific-sentiment tasks is extraction of aspect-specific words and opinion-specific words. In particular, a general framework of summarizing reviews of a certain product is to first identify different aspects of the given product and then extract specific opinion expressions for each aspect. For example, aspects of a restaurant may include *food*, *staff*, *ambience*, and *price*, and opinion expressions for staff may include *friendly*, *rude*, etc. A separation of aspect and opinion words can be very useful. Aspect-specific opinion words can be used to construct a domain-dependent sentiment lexicon and applied to tasks such as sentiment classification. They can also provide more informative descriptions of the product or service being reviewed. For example: two example review sentences from the restaurant domain:

The food was tasty.

The waiter was quite friendly.

We can see that there is a strong association of *tasty* with *food* and similarly of *friendly* with *waiter*. While both *tasty* and *friendly* are specific to the restaurant domain, they are each associated with only a single aspect, namely *food* and *staff*, respectively. Besides these aspect-specific opinion words, there are also general opinion words such as great in the sentence "*The food was great!*" These general opinion words are shared across aspects, as opposed to aspect-specific opinion words that are used most commonly with their corresponding aspects.

Zhao et al. [Zhao et al., 2010] introduce a general opinion model and T aspect-specific opinion models to capture these different opinion words. Their model is an extension of LDA, but captures both aspect words and opinion words.

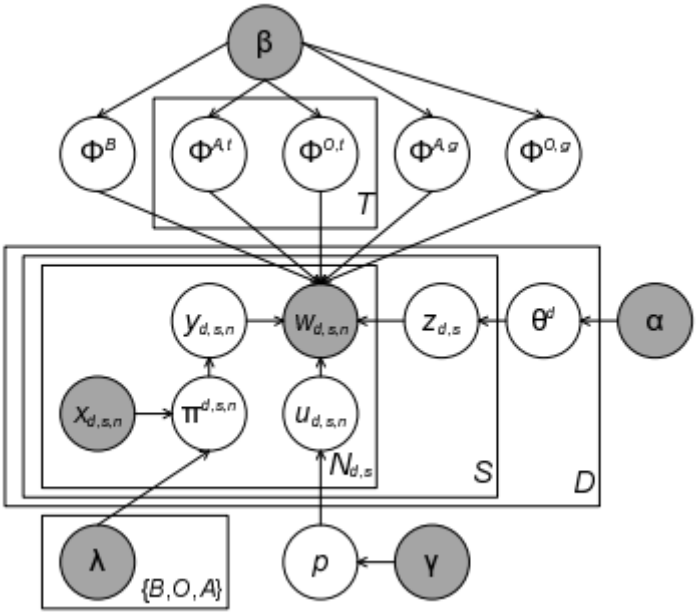

Figure 6: MaxEnt-LDA model

Zhao et al's model first draws several multinomial word distributions from a symmetric Dirichlet prior with parameter β: a background model ϕ^{B}, a general aspect model ϕ^{Ag}, a general opinion model ϕ^{Og}, T aspect models $\{\phi^{At}\}_{t=1}^{T}$, and T aspect-specific opinion models $\{\phi^{Ot}\}_{t=1}^{T}$. All these are multinomial distributions over the vocabulary, which they assume has V words. Then for each review document d, they draw a topic distribution $\theta^{d} \sim Dir(\alpha)$ as in standard LDA. For each sentence s in document d, they draw an aspect assignment $z_{d,s} \sim Multi(\theta^{d})$.

Now for each word in sentence s of document d, they have several choices: The word may describe the specific aspect, or a general aspect, or an opinion either specific to the aspect or generic, or a commonly used background word. To distinguish these choices, they introduce two indicator variables, $y_{d,s,n}$ and $u_{d,s,n}$ for the nth word $w_{d,s,n}$.

They draw $y_{d,s,n}$ from a multinomial distribution over {0, 1, 2}, parameterized by $\pi^{d,s,n}$. $y_{d,s,n}$ determines whether $w_{d,s,n}$ is a background word, aspect word, or opinion word. They set $\pi^{d,s,n}$ using a maximum entropy (MaxEnt) model applied to a feature vector $x_{d,s,n}$ associated with $w_{d,s,n}$. $x_{d,s,n}$ includes features: {previous, the current and the next words} and POS tag features {*POS*i−1, *POS*i, *POS*i+1}. They draw $u_{d,s,n}$ from a Bernoulli distribution over {0, 1} parameterized by p, which in turn is drawn from a symmetric *Beta(γ).*

Datasets: a restaurant review data set previously used in Ganu et al. [Ganu et al., 2009] and Brody and Elhadad [Brody and Elhadad, 2010], and a hotel review data set previously used in Baccianella et al. [Baccianella et al., 2009].

Results:

a) Aspects identification:

Table 8: Results of aspects identification on restaurant.

Aspect	Method	Precision	Recall	F-1
Staff	LocLDA	0.804	0.585	0.677
	ME-LDA	0.779	0.540	0.638
Food	LocLDA	0.898	0.648	0.753
	ME-LDA	0.874	0.787	0.828
Ambience	LocLDA	0.603	0.677	0.638
	ME-LDA	0.773	0.558	0.648

They compared aspect identification with Local-LDA [Brody and Elhadad, 2010] and reported better accuracy in some aspects.

b) Opinion identification:
They quantitatively evaluated the quality of the aspect specific opinion words identified by ME-LDA and compared it to their base lines. BL-2 is based on LDA and most frequent adjectives were chosen as aspect-specific opinion words.

Table 9: Average P@n of aspect-specific opinion words on restaurant.

Method	P@5	P@10	P@20
ME-LDA	0.825	0.700	0.569
BL-1	0.400	0.450	0.469
BL-2	0.725	0.650	0.563

Sauper et al. [Sauper et al., 2010] present a different approach from those above. Their representation of context encodes more than the relevance-based binary distinction considered in past work. Their algorithm adjusts the content model dynamically for a given task rather than pre-specifying it. Their method is fully unsupervised.
Model: First the document-level HMM generates a hidden content topic sequence T for the sentences of a document. This content component is parameterized by θ and decomposes in the standard HMM fashion. Then the label sequences for each sentence in the document are independently modeled as CRFs, which condition on both the sentence features and the sentence topic.

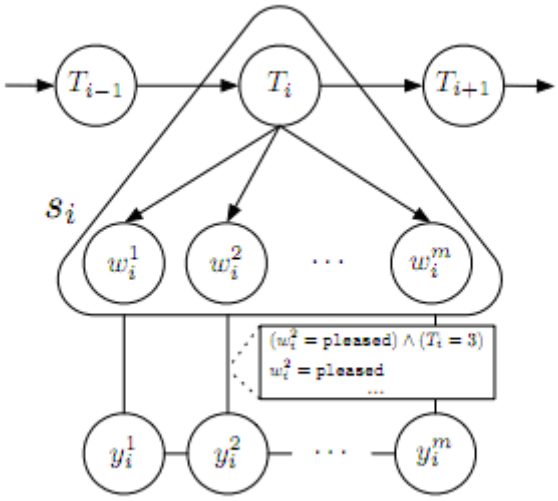

Figure 7: A graphical depiction of our model for sequence labeling tasks. The *T*i variable represents the content model topic for the *i*th sentence s_i. The words of s_i, $(w_i^1,...,w_i^m)$ each have a task label $(y_i^1,...,y_i^m)$. Note that each token label has an undirected edge to a factor containing the words of the current sentence, s_i as well as the topic of the current sentence *T*i.

Dataset: DVD reviews from the website IGN.com
Results: Sauper et al. compared their method against a simplified variant of their method wherein a content model is induced in isolation rather than learned jointly in the context of the underlying task. They refer to the two methods as the No Content Model (*NoCM*) and Independent Content Model (*IndepCM*) settings, respectively. The Joint Content Model (*JointCM*) setting refers to their full model, where content and task components are learned jointly.

Table 10: The error rate on the multi-aspect sentiment ranking. They reported mean L1 and L2 between system prediction and true values over all aspects.

	L_1	L_2
NoCM	1.37	3.15
IndepCM	1.28†*	2.80†*
JointCM	**1.25**†	**2.65**†*
Gold	1.18†*	2.48†*

They reported that the joint content and sentiment model gets better results than the no content or independent content models.

Summary
The following summarizes the reported performance of previous work on similar datasets.

Aspect extraction comparison of reported models:

ME_LDA [Zhao et al., 2010] > (reported better results) LocLDA [Brody and Elhadad, 2010] > LDA
SDWP [Zhan and Li, 2011] > LDA
Constrained_LDA [Zhai et al., 2011] > LDA
AEOT [Zhang et al., 2011] > LDA

All the above models are unsupervised models and were compared with a pure LDA baseline. There was no comparison between supervised and unsupervised models in aspect extraction.

Sentiment analysis comparison:

MG_LDA [Zhao et al., 2010] > LDA
ASUM [Jo and Oh, 2011] > JST [Lin and He, 2009] < supervised classifiers
LRR [Wang et al., 2010] < supervised classifiers
Word Vectors [Maas et al., 2011] > LDA
JointCM [Sauper et al., 2010] > 2-phase approach [Sauper et al., 2010] > NoCM [Sauper et al., 2010]

The unsupervised models above were compared with an LDA baseline and some of them were compared with supervised classifiers and were found less effective than supervised classifiers.

In this research, we review two aspects not previously investigated:

- Do topic models help in supervised aspect identification and sentiment detection?
- We want to compare results across multiple datasets that have been used in previous work but not previously compared.

Chapter 5: Proposed Method

Use of Topics Distributions as Features for Aspects Detection and Sentiment Classification

5.1. Methodology

We saw in previous research the use of supervised and unsupervised approaches. There are two major approaches of aspect-grained sentiment analysis:

- Two-phase approaches, where the first phase attempts to extract the aspects of an object that users frequently rate and the second phase classifies and aggregates sentiment over each of these aspects.
- Joint model of a content model (aspects discovery) and a sentiment analysis that discovers aspects and sentiment simultaneously.

We propose a two-step semi-supervised novel approach for aspect sentiment classification. Following linguistic intuition and previous research, rich contextual information can benefit text analysis applications such as aspects extraction and sentiment classification. The induced content structure is learned from a large un-annotated corpus. We demonstrate that exploiting content structure yields significant improvements over approaches that rely only on local context. We propose a combination of words appearance with LDA topics as features in supervised learning.

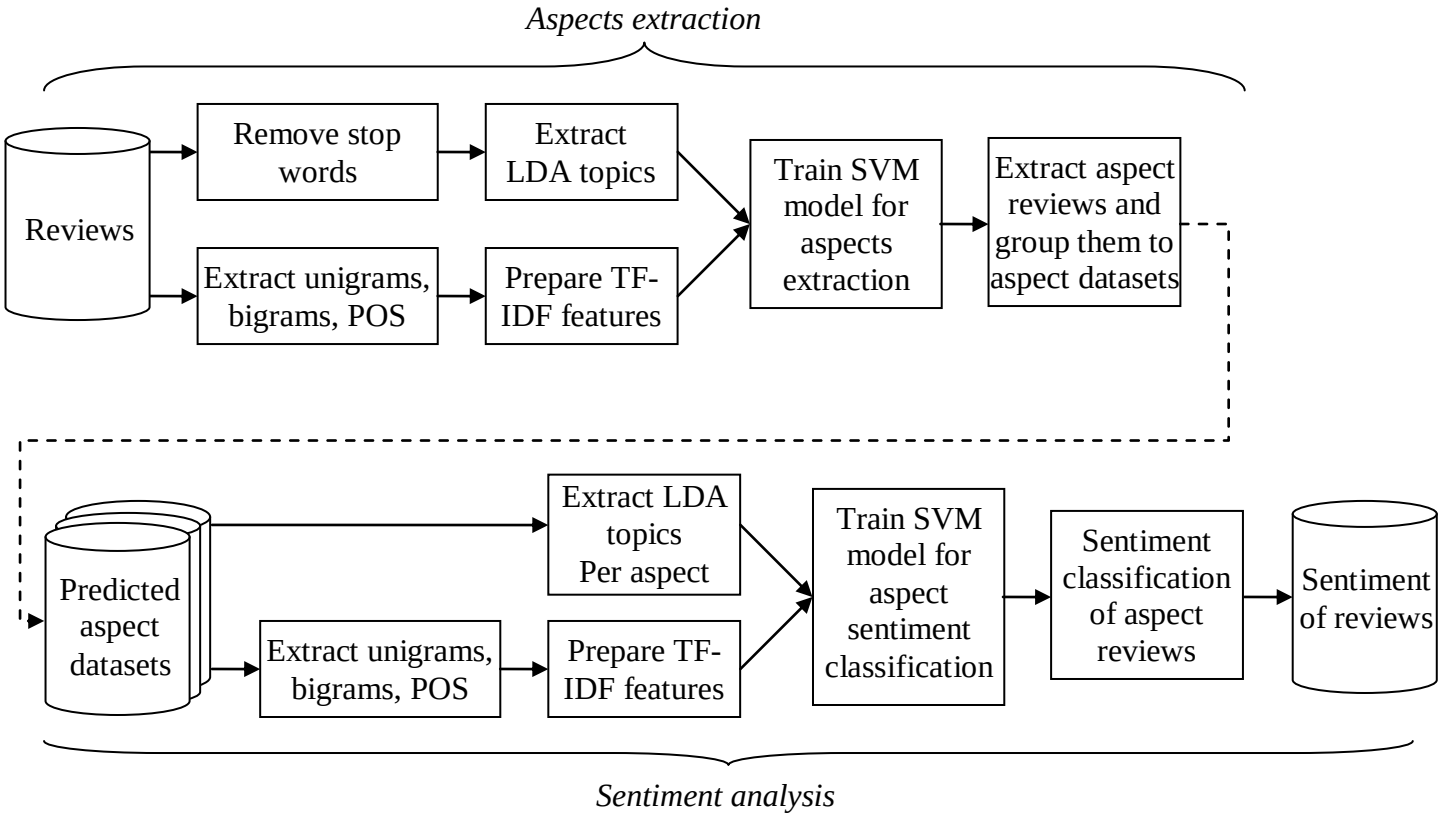

Figure 8: Flow chart describes major steps of our approach.

General Overview of the Proposed Method

We propose a *two-phase* aspect sentiment detection approach which is graphically described in Figure 8. The *first phase* attempts to extract the aspects of an object. The *second phase* classifies sentiment over each of these aspects. Our approach thus extends on the two-phase approaches surveyed in Section 4.2. Our contribution is to combine unsupervised techniques for aspect and sentiment detection to provide features that drive a context-aware supervised model of aspect-sentiment classification.

As reviewed in Section 4.2, most aspect detection methods rely on unsupervised (or weakly semi-supervised) topic models: Multi-Grain LDA [Titov and McDonald, 2008], the sentence-based topic classification method of [Brody and Elhadad, 2010], Word-level Aspect Extraction [Maas et al, 2011], the Author-Experience-Object-Topic model [Zhang et al, 2011], constrained LDA [Zhai et al, 2011], and Semantic-Dependent Word Pairs Generative Model [Zhan and Li, 2011].

In order to construct better classifiers for aspect and sentiment classification, we propose to use *unsupervised* features along with bag-of-words features. Our unsupervised features are topics distribution over inferred aspects for each sentence in the data. In order to infer the salient aspects in the data, we specifically use topics distributions inferred by a local version of LDA in a manner similar to [Brody and Elhadad, 2010].

We further construct a supervised classifier (SVM) in order to build a sentiment classifier per aspect. Most previous research used supervised classification methods with different kind of features (as reviewed in Section 4.2.2): document level bag of word and n-gram features [Pang et al., 2002], with feature selection based on subjectivity filtering [Pang and Lee, 2004], or syntactic features [Matsumoto et al., 2005], [Wilson et al., 2005], [Eriksson, 2005] and [Green and Resnik, 2009]. In all cases, it was found that a bag of words features with some form of feature selection provides a solid baseline.

Our contribution is to complement such a bag of words feature set with an aspect-specific context representation encoded in the form of topic labels. The SVM classifier for sentiment classification is trained using bag of words features together with topic labels inferred over aspect classified sentences. Overall, this stage of our method is an aspect-aware extension of [Pang et al 2002] with filtering by aspect and the addition of LDA topics as features.

The overall method adopts the following process:

Input: Reviews datasets with aspects and aspect-sentiment tags for each aspect-sentiment review. We processed all datasets which were described in Section 3.1.

Output: Discovered aspects and their sentiments per each review of the input datasets.

Processing: The steps of the method as shown in Figure 8 are:

- **Aspect Extraction**:
  - Prepare *unsupervised* features per each review:
    - Remove stop words from all reviews.

        - Extract LDA topics for each review using a standard implementation of LDA [McCallum and Kachites] over the whole dataset.
    - Prepare bag-of-words features for each review:
        - Extract bag-of-words features: unigrams and part of speech features from the input dataset.
        - Calculate TF-IDF weight for each feature.
    - Train an SVM model for each aspect based on (bag of words, parts of speech tf-idf, LDA-topics) as features. We use 10-fold cross-validation technique to validate the accuracy of our model.
    - Extract aspect predicted reviews and group them into aspect datasets (one dataset per aspect).

- **Sentiment Analysis**:
    - For each aspect dataset:
        - Train an aspect-specific LDA topic model
        - Prepare bag-of-words features for each review:
            - Extract bag-of-words features: unigrams and part of speech features from aspect dataset.
            - Calculate TF-IDF weight for each feature.
        - Train an SVM model for each sentiment based on (bag of words, parts of speech, tf-idf, LDA-topic) as features. We used 10-fold cross-validation technique to validate the accuracy of our model.
        - Perform sentiment classification of aspect reviews.

In the following sections, we describe in details each step of our approach with examples from various review datasets.

5.2. Aspect Classification

In order to explore aspects in the reviews we used a supervised learning method with unsupervised features. We used Support Vector Machine, where vector features are constructed from unigrams, bi-grams, and distribution over inferred aspects for each sentence in the data. In order to infer the salient aspects in the data, we use topics distributions inferred by a local version of LDA [Brody and Elhadad, 2010].

Latent Dirichlet Allocation

Latent Dirichlet allocation (LDA) is a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document.

LDA assumes the following generative process for each document w in a corpus D:

1. Choose $N \sim$ Poisson(ξ).
2. Choose $\theta \sim$ Dir(α).
3. For each of the N words w_n:
   (a) Choose a topic $z_n \sim$Multinomial(θ).
   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

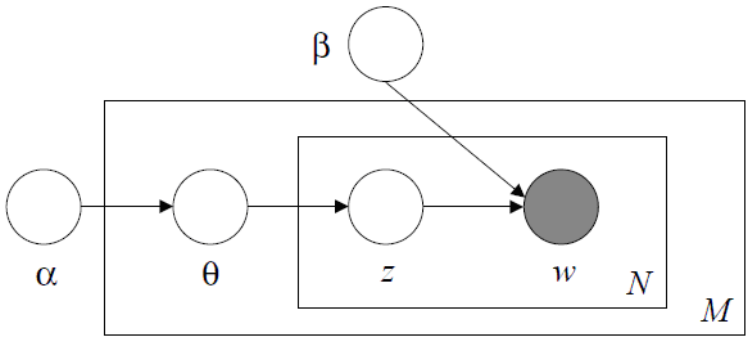

Figure 9: Graphical model representation of LDA. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

The LDA model is represented as a probabilistic graphical model in Figure 9. As the figure makes clear, there are three levels to the LDA representation. The parameters α and β are corpus-level parameters, assumed to be sampled once in the process of generating a corpus. The variables θ_d are document-level variables, sampled once per document. Finally, the variables z_{dn} and w_{dn} are word-level variables and are sampled once for each word in each document.

Local Version of LDA

We present a local version of LDA introduced in [Brody and Elhadad, 2010]. According to previous research, LDA is not suited to the task of aspect detection in reviews, because it tends to capture global topics in the data, rather than ratable aspects relevant to the review. In order to prevent the inference of global topics and direct the model towards ratable aspects, we treated each sentence as a separate document. The output of the model

is a distribution over inferred aspects for each sentence in the data. Stop-words were removed from input to LDA in order to filter out generally common words in English, which do not carry any valuable information.

For example:

Table 11: aspects inferred for the restaurant domain [Ganu et al., 2009]

Aspect index	Inferred aspect	Representative words
0	Staff	Table, wait, waiter, order, seated, minutes, waitress, reservation, asked, check, hour, manager, reservations, waiting, hostess
1	Location	place, great, love, nice, perfect, fun date spot, neighborhood, live, happy, work, street location, park, cute café, stop review
2	Anecdotes	dinner time, night, friends, worth, friend, group, party, birthday, evening, brunch
3	Value	great food experience, amazing, wonderful prices, delicious, fantastic, thought, absolutely cheap, reasonable, coffee, loved, awesome, dining ambience
4	Food – General	Dishes, fresh, ordered, fish, tasty, dish, portions, delicious, served, appetizer, taste, small, entree, seafood, pasta, main plate
5	Food-General	menu, sushi, city, restaurants, NYC, places, favorite, Italian, York, French, high, authentic Thai cuisine, Indian, Manhattan, simply, NY ,tasting
6	Ambience/Mood	Atmosphere, decor, room, dining, nice, music, feel, romantic, cool, scene, warm, space, beautiful, crowd, ambience, cozy, makes, loud, comfortable
7	General	restaurant, recommend, special, highly, lunch, 'd chef day, recommended, real, week, burger, house, wife, husband, business, deal, choice
9	Physical Atmosphere	bar, people, make, restaurant, time, big, tables, small, area, large, lot, money, kitchen, sit, crowded, long, seating, door, kind
10	Wine & Drinks	Good, food wine, drinks, price, pretty, list, quality, average, excellent, expensive, bit, selection, glass, fine, bottle
11	Service	service, staff, friendly, food, excellent, attentive, rude, reviews, extremely, slow waiters, owner, terrible, pleasant attitude, surprised, horrible server
12	Bakery	Dessert, pizza, chocolate, hot desserts, cold, tasted, couple, home, cake, worst, cream, eaten, world, tea
13	Main Dishes	Chicken, steak, cheese, salad, sauce, shrimp, bread, meat, tuna, sweet, soup, fries, fried, lobster, pork, duck, salmon, rice, beef

Following are examples of topics distribution for each sentence.

Review 1: *"He argues with me, realizes his mistake then retrieves my order."*
Its aspects distribution is:

Table 12: Topics distribution of review 1.

Topic:	11	13	12	10	9	8	7	6	5	4	3	2	1	0
Weight:	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

This sentence describes a positive review about staff. We can see that aspect number 11 (service) got a highest score over all other aspects.

Review 2:*"the menu claimed the bagel was jumbo-sized and toasted and it was neither--small and cold"*
Its aspects distribution is:

Table 13: Topics distribution of review 2.

Topic:	12	10	5	4	2	1	13	11	9	8	7	6	3	0
Weight:	0.28	0.14	0.14	0.14	0.14	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

This sentence describes a negative opinion about food. We can see that aspect number 12 (bakery) got a highest score over all other aspects.

Review 3: *"The burger is fantastic as are the salads the fries the bloody marys and the atmosphere which harkens back to a better new york"*
Its aspects distribution is:

Table 14: Topics distribution of review 3.

Topic:	13	2	12	5	11	10	9	8	7	6	4	3	1	0
Weight:	0.42	0.28	0.14	0.14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

This sentence describes a positive opinion about food and atmosphere. We can see that aspect number 13 (Bakery) and 5 (Food – General) got higher scores than all other aspects.

From the following example we can see that topics distribution over a sentence can be valuable features in aspects classification.

Combination of unsupervised features (LDA topics) in SVM

Above, we described LDA topics as part of the features in SVM learning and testing. Along with LDA topics features, we used additional features:

- **Unigrams**: Different aspects are described by different aspects-specific words. For example, the Food aspect in the restaurant domain is described by food-specific words such as: *chicken*, *steak*, *cheese*, *salad*, *sauce*, *shrimp*, *bread*. While the Service aspect is described by *service*, *staff*, *friendly*, *food*, *excellent*, *attentive*, *rude*, *reviews*, *extremely*, *slow*, *waiters*. We use all unigrams, despite the fact that not all unigrams describe aspects, because in previous research it was shown that additional information increased accuracy.
- **Bi-grams**: bi-grams can be useful in capturing aspect phrases such as *battery life*, while the unigram *life* alone is a general word.
- **Part-of-speech**: Part-of-speech (POS) information is commonly exploited in sentiment analysis and opinion mining. Adjectives are good indicators of sentiment, and have sometimes been used to guide feature selection for sentiment classification. Aspect words and opinion words usually play different syntactic roles in a sentence. Aspect words tend to be nouns while opinion words tend to be adjectives. Their contexts in sentences can also be different. But we do not want

to use strict rules to separate aspect and opinion words because there are also exceptions. We used POS as part of vector features.

SVM:

Support vector machines (SVMs) have been shown to be highly effective at traditional text categorization. The basic idea behind the training procedure is to find a hyperplane, represented by vector $\vec{w}_i$, which not only separates the document vectors in one class from those in the other, but for which the separation, or margin, is as large as possible. This search corresponds to a constrained optimization problem; letting $c_j \in \{1, -1\}$ (corresponding to positive and negative) be the correct class of document d_j, the solution can be written as:

$$\vec{w}_i := \sum_j \alpha_j c_j \vec{d}_j, \alpha_j \geq 0,$$

where α_j are obtained by solving a dual optimization problem. Those $\vec{d}_j$ such that α_j is greater than zero are called *support vectors*, since they are the only document vectors contributing to $\vec{w}$. Classification of test instances consists simply of determining which side of $\vec{w}$'s hyperplane they fall on.

Vector feature weights

We used a tf.idf weighting scheme rather binary weights.

The classic *tf.idf* formula assigns weight w_i to term i in document D as:

$$w_i = tf_i * idf_i = tf_i * \log \frac{N}{df_i}$$

where tfi is the number of times term i occurs in D, $idfi$ is the inverse document frequency of term i, N is the total number of documents, and dfi is the number of documents that contain term i. We use tf-idf because it captures corpus-wide importance of words; words that are more frequent in a document than expected across all documents are more relevant than words that are frequent across all documents.

5.3. Sentiment Classification

5.3.1. Overall rating against aspect-specific rating

Interactions between aspect and sentiment play an important role in opinion mining. Different aspects in the same review can contain different sentiment ranking. For example:

> *"The food was okay, not great, not bad. Our favorite part, though, was the show!"*

A user not interested in ambience would probably not want to dine at this restaurant. However, a recommendation based on star ratings would label this restaurant as a high-quality restaurant. This fact will damage an overall review prediction.

If all but the main topic can be disregarded, then one possibility is as follows: simply consider the overall sentiment detected within the document – regardless of the fact that it may be formed from a mixture of opinions on different aspects to be associated with the primary topic, leaving the sentiment towards other topics undetermined. Later, we will

show that an overall review sentiment classifier shows lower results than aspect-specific results in some datasets.

5.3.2. Features

We used the same features as in aspects extraction: unigrams, bigrams, and topics distribution over each sentence.

- LDA topics distribution:
  Topics are useful latent structures to explain semantic association. Latent topics are thus discovered by identifying groups of words in the corpus that frequently occur together within documents. Therefore, with a large enough number of topics, we can discover topics that describe specific aspects sentiments of reviews.

For example:

Table 15: Discovered topics in the Hotels data base.

Aspect	**ID**	**Representative words**
Cleanliness	29	rude, told, asked desk, bad, terrible, worst night, moved finally, awful, working tiny man, money, checked, manager complained
Cleanliness	49	walls, toilet dirty, wall dark, bad smell, carpet, paper, poor, worst, worn, tiny, terrible, horrible, shabby work, worse, dated
Cleanliness	37	room, breakfast good, small, clean, nice room staff, average, walk, night, great, decent, large, fine, quiet, noise, single suite
Cleanliness	6	breakfast room staff good, helpful, walk, great, quiet, excellent room, comfortable, clean restaurants, large, friendly, spacious, position, Jacuzzi

Topics 29 and 49 describe cleanliness aspects with negative sentiment.
Topics 37 and 6 have a strong positive sentiment.

The following review describes cleanliness with negative sentiment:

"overpriced need manager best staff floor tiny good unhelpful finally room better new lovely smell star stars unfriendly desk man room eventually barely"

(In this database, we do not have full sentences, only unigrams features.)
:

Table 16: Topics distribution of this review.

Topic:	29	58	49	4	1	47	15	14	10	2
Weight:	0.31	0.10	0.10	0.10	0.10	0.05	0.05	0.05	0.05	0.05

We can see that topic number 29 got the highest weight.

The same with this review*:*

"saying stains staff unhelpful carpet stay phone room"

which got:

Table 17: Topics distribution of this review.

Topic:	29	58	46	14	53	52	49	36	4	3	0
Weight:	0.22	0.16	0.11	0.11	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Review:

"toilet stained wouldn wasn old staff friendly dirty good walk charged helpful great room quiet room small stank carpets"

Got:

Table 18: Topics distribution of this review.

Topic:	49	47	40	38	26	32	29	27	10	6
Weight:	0.27	0.11	0.11	0.11	0.11	0.05	0.05	0.05	0.05	0.05

Here, topic 49 got the highest weight and it represents negative sentiment about cleanliness aspect.

Positive sentiment examples:

Review:

"highly wasn breakfast fine staff city better recommended friendly price paid buffet great helpful good room clean night"

Got:

Table 19: Topics distribution of this review.

Topic:	37	54	34	20	48	47	41	26	25	18	16
Weight:	0.23	0.11	0.11	0.11	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Review:

"terrace restaurants spacious good new large spotless steps definitely great disappointment spanish given tiny door room comfortable wonderful breakfast buffet stay room clean "

Got:

Table 20: Topics distribution of this review.

Topic:	6	43	41	25	58	56	54	46	38	31	30	23	13	10
Weight:	0.2	0.1	0.1	0.1	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05

Topics 6 and 37 describe a positive opinion about the cleanliness of a hotel.

Summary

Our method applies supervised learning using topic-model features in addition to lexical features. We use a 2-step method (not a joint-model) of first categorizing sentences into aspects, then classifying each sentence into sentiment given the predicted aspect of the sentence. Data observation indicates a strong correlation between aspect, topic and sentiment, thus supporting intuitively the strategy of learning topic models per aspect.

Chapter 6: Results

6.1. Datasets Summary

Table 21: We worked with the following datasets (as described in Chapter 3).

Dataset	Number of aspects	Number of sentences
Restaurants	6	80,000
Hotels	7	49,471
Multi-Domain	4	3,684
DVD	4	2,660

6.2. Unbalanced Data Sets

Our datasets are characterized by a strong unbalance: the ratio of sentiment sentences to overall sentences, and sentences for each aspect vs. overall sentences is very small. Dealing with unbalanced datasets requires some adjustment to classifiers to avoid introducing systematic bias towards negative classification.

Classic approaches proposed for tackling the class imbalance problem include:

1. Upsizing the small class at random,
2. Upsizing the small class at “focused" random,
3. Downsizing the large class at random,
4. Downsizing the large class at “focused" random,, and
5. Altering the relative costs of misclassifying the small and the large classes.

Methods 1 and 2 consist of re-sampling patterns of the small class (either completely randomly or randomly but within parts of the input space close to the boundaries with the other class) until the quantity data from the small class is the same as from the large one. Methods 3 and 4 consist of eliminating data from the large class (either completely randomly or randomly but within parts of the input space far away from the boundaries with the large class) until both classes have equal amounts of data. Finally, method 5 consists of reducing the relative misclassification cost of the large class (or, equivalently, increasing that of the small one) to make it correspond to the size of the small class.

From previous research [Japkowicz and Stephen, 2002], the last method, cost-modifying, was found to be more effective than others. We compared two methods: random under-sampling and cost factor:

1) Random under-sampling:
   Random under-sampling of a large class until it reaches the size of a small class. This under-sampling is randomly performed 10 times and average results are reported.
2) Cost factor:
   We used an implementation of the cost factor in SVMlight (-j argument, described in [Morik et al., 1999]): These authors introduce cost factors C+ and C- to be able to adjust the cost of false positives vs. false negatives. Without loss of generality, the cost for an FP is always 1. The cost for an FN is usually suggested to be the ratio of negative samples over positive samples. We reported two cost functions:

   $$\text{Cost function 1} = \frac{number\_of\_review\_in\_small\_class}{number\_of\_review\_in\_large\_class}$$

Cost function 2 = inverse of cost function 1

For the baseline, we took the data set as is without any sampling and any cost factor. We ran a comparison of the methods on all our data sets and took the method that showed the best results. For example, the hotel data set aspects extraction:

Table 22: The hotel data set aspects extraction Accuracy, Precision and Recall.

Aspect	*Baseline*	*Cost function 1*	*Cost function 2*	*Random under-sampling*
Service	A = 87.25 P = 84.08 R = 91.76	A = 86.81 P = 83.16 R = 92.23	A = **87.33** P = **84.14** R = **91.96**	A = 87.07 P = 83.73 R = 91.93
BService	A = **96.54** P = **98.72** R = **94.30**	A = 96.22 P = 98.37 R = 94.02	A = 96.37 P = 98.47 R = 94.21	A = 96.43 P = 98.71 R = 94.11
Checkin	A = 89.33 P = 88.34 R = 91.85	A = 89.17 P = 88.28 R = 91.52	A = **89.93** P = **89.05** R = **92.01**	A = 89.56 P = 88.65 R = 91.82
Value	A = **91.27** P = **89.64** R = **93.58**	A = 91.16 P = 89.47 R = 93.51	A = 91.12 P = 89.41 R = 93.49	A = 91.20 P = 89.53 R = 93.50
Rooms	A = 89.88 P = 87.25 R = 93.79	A = 89.81 P = 87.26 R = 93.60	A = **90.09** P = **87.53** R = **93.80**	A = 90.06 P = 87.68 R = 93.57
Clean	A = 92.41 P = 87.93 R = 98.60	A = **92.53** P = **88.46** R = **98.16**	A = 92.21 P = 87.88 R = 98.14	A = 92.34 P = 88.01 R = 98.36
Location	A = 97.67 P = 96.14 R = 99.38	A = **97.72** P = **96.07** R = **99.54**	A = 97.50 P = 95.77 R = 99.44	A = 97.60 P = 95.81 R = 99.61

In this dataset, results are pretty close to each other. Best results are appears in bold font. We choose to run with baseline in aspects detection and cost function 1 in sentiment analysis as best practice.

6.3. Aspects Extraction

We used SVM-light [Joachims, 2008] implementation of SVM with default parameters. We used this binary classifier in order to predict multi-class in a one-versus-all model:

1. Build a classifier for each class, where the training set consists of the set of documents in the class (positive labels) and its complement (negative labels).
2. Given the test set, apply each classifier separately. The decision of one classifier has no influence on the decisions of the other classifiers.

We used a standard implementation of LDA [McCallum and Kachites]. The parameters we employed were standard, out-of-the-box settings ($\alpha = 0.1$, $\beta = 0.1$, 2000 iterations).

6.3.1. Hotels Dataset

Table 23: The accuracy of aspect detection on the Hotel dataset using SVM with unigram and topic features.

Aspect	*Baseline*	*15 topics*	*20 topics*	*30 topics*	*40 topics*	*60 topics*	*100 topics*
Service	A = 89.06 P = 87.38 R = 92.19 F1 = 89.72	A = 88.96 P = 87.63 R = 92.08 F1 = 89.80	A = 88.96 P = 87.77 R = 92.01 F1 = 89.84	A = 89.18 P = 87.96 R = 92.13 F1 = 90.00	A = 89.20 P = 88.08 R = 92.00 F1 = 90.00	A = **89.47** P = 88.17 R = 92.34 F1 = **90.20**	A = 89.41 P = 88.48 R = 91.70 F1 = 90.00
BService	A = 96.98 P = 98.83 R = 95.14 F1 = 96.95	A = 97.24 P = 98.24 R = 96.26 F1 = 96.95	A = 97.14 P = 98.32 R = 96.04 F1 = 97.16	A = 97.32 P = 98.60 R = 96.07 F1 = 97.32	A = 97.51 P = 98.69 R = 96.36 F1 = 97.51	A = 97.66 P = 98.96 R = 96.36 F1 = 97.64	A = **97.69** P = 99.28 R = 96.11 F1 = **97.67**
Checkin	A = 92.35 P = 91.88 R = 93.25 F1 = 92.56	A = 92.46 P = 92.04 R = 93.23 F1 = 92.63	A = 91.95 P = 91.49 R = 92.87 F1 = 92.17	A = 92.10 P = 91.38 R = 93.25 F1 = 92.31	A = 92.38 P = 91.82 R = 93.33 F1 = 92.57	A = 92.66 P = 92.46 R = 93.12 F1 = 92.79	A = **93.28** P = 92.79 R = 94.04 F1 = **93.40**
Value	A = 91.82 P = 89.83 R = 95.14 F1 = 92.41	A = 91.83 P = 89.84 R = 95.29 F1 = 92.48	A = 91.86 P = 89.81 R = 95.41 F1 = 92.52	A = 91.62 P = 89.59 R = 95.23 F1 = 92.32	A = 91.85 P = 89.88 R = 95.28 F1 = 92.50	A = 91.85 P = 90.06 R = 94.94 F1 = 92.43	A = **92.14** P = 90.32 R = 95.25 F1 = **92.72**
Rooms	A = 92.90 P = 89.22 R = 98.63 F1 = 93.69	A = 92.73 P = 89.24 R = 98.31 F1 = 93.55	A = 93.08 P = 90.00 R = 98.04 F1 = 93.85	A = 93.01 P = 89.84 R = 98.09 F1 = 93.78	A = 93.07 P = 90.1 R = 97.91 F1 = 93.84	A = 93.09 P = 90.01 R = 98.04 F1 = 93.85	A = **93.18** P = 90.07 R = 98.14 F1 = **93.93**
Clean	A = 94.27 P = 91.30 R = 98.49 F1 = 94.76	A = 94.38 P = 91.62 R = 98.37 F1 = 94.87	A = 94.39 P = 91.81 R = 98.14 F1 = 94.87	A = **94.86** P = 92.28 R = 98.52 F1 = **95.30**	A = 94.71 P = 92.26 R = 98.27 F1 = 95.17	A = 94.83 P = 92.34 R = 98.37 F1 = 95.26	A = 94.73 P = 92.21 R = 98.29 F1 = 95.15
Location	A = 97.99 P = 96.56 R = 99.56 F1 = 98.03	A = 98.01 P = 96.57 R = 99.58 F1 = 98.05	A = 97.95 P = 96.74 R = 99.27 F1 = 97.99	A = 97.97 P = 96.69 R = 99.37 F1 = 98.01	A = 97.94 P = 96.59 R = 99.42 F1 = 97.99	A = 97.93 P = 96.59 R = 99.40 F1 = 97.98	A = **98.02** P = 97.06 R = 99.067 F1 = **98.05**

In this dataset we used unigram features for baseline and 15, 20, 30, 40, 60, and 100 topic distribution features. Use of topic distribution features improved accuracy in all aspect detections. For instance, check-in aspect detection improved from 92.35% to 93.28%, and F1-score from 92.56% to 93.40% using 100 topics distribution.

The best results are obtained for K=100 topics with a significant improvement in accuracy and F1 measure for all aspects.

6.3.2. DVD dataset

Table 24: The accuracy of aspect detection on the DVD dataset.

Aspect	*no topics*	*10 topics*	*20 topics*	*100 topics*
Audio	A = 98.88 P = 99.69 R = 95.84	A = 99.04 P = 98.93 R = 97.22	A = **99.04** P = **99.23** R = **96.92**	A = 99.04 P = 99.53 R = 96.61
Extras	A = 94.96 P = 94.75 R = 84.76	A = 95.0 P = 93.21 R = 86.61	A = **95.38** P = **94.60** R = **86.77**	A = 95.34 P = 95.23 R = 85.84
Movie	A = 93.60 P = 84.33 R = 92.77	A = 94.02 P = 85.12 R = 93.38	A = 93.98 P = 85.18 R = 93.38	A = **94.33** P = **89.90** R = **87.69**
Video	A = 97.96 P = 99.04 R = 92.76	A = 98.11 P = 98.13 R = 94.30	A = **98.27** P = **98.72** R = **94.31**	A = 98.11 P = 99.20 R = 93.23

In this data set, topic features also improved aspect detection results.

6.3.3. Restaurants dataset

Table 25: The accuracy of aspect detection on the restaurants dataset.

Aspect	*No topics*	*10 topics*	*14 topics*	*20 topics*	*100 topics*	*200 topics*
Food	A = **95.00** P = **94.44** R = **95.64** F1 = **95.00**	A = 94.40 P = 93.76 R = 95.14 F1 = 94.40	A = 94.47 P = 93.87	A = 94.48 P = 93.82 R = 95.23 F1 = 94.52	A = 94.53 P = 93.93 R = 95.21 F1 = 94.50	
Ambience	A = **95.02** P = **94.50** R = **95.61** F1 = **95.00**		A = 94.66 P = 93.70 R = 95.79 F1 = 94.73		A = 94.53 P = 93.86 R = 95.32 F1 = 94.58	A = 94.72 P = 93.78 R = 95.80 F1 = 94.70
Staff	A = **97.53** P = **97.88** R = **99.16** F1 = **98.51**	A = 97.31 P = 97.69 R = 99.08 F1 = 98.38	A = 97.31 P = 97.70 R = 99.07 F1 = 98.38	A = 97.34 P = 97.72 R = 99.08 F1 = 98.40	A = 97.36 P = 97.73 R = 99.09 F1 = 98.41	A = 97.37 P = 97.76 R = 99.08 F1 = 98.42
Price	A = **98.77** P = **99.33** R = **99.34** F1 = **99.33**	A = 98.64 P = 99.26 R = 99.27 F1 = 99.27	A = 98.66 P = 99.26 R = 99.29 F1 = 99.28	A = 98.69 P = 99.29 R = 99.29 F1 = 99.29	A = 98.68 P = 99.27, R = 99.30 F1 = 99.29	A = 98.69 P = 99.26 R = 99.33 F1 = 99.29

In this dataset, no significant improvement was made by use of topic distribution, probably due to the homogeneous nature of the data of this data set.

6.3.4. Multi-domain Dataset

Table 26: The accuracy of aspect detection on the multi-domain dataset.

Product type	*No topics*	*10 topics*	*14 topics*	*20 topics*	*30 topics*	*50 topics*	*100 topics*
Books	A = 94.28 P = 95.31 R = 93.33	A = 94.76 P = 95.33 R = 94.40	A = 95.11 P = 95.29 R = 95.24	A = 95.47 P = **95.75** R = 95.36	A = 95.35 P = 95.45 R = 95.47	A = **95.47** P = 95.64 R = **95.48**	A = 94.93 P = 95.21 R = 94.88
DVD	A = 91.73 P = 90.23 R = **94.83**	A = 91.00 P = 92.06 R = 90.33	A = 91.83 P = 93.49 R = 90.33	A = 92.66 P = **94.50** R = 90.33	A = 92.00 P = 94.90 R = 88.99	A = **92.83** P = 94.64 R = 91.00	A = 91.83 P = 94.49 R = 89.00
Electronics	A = 91.73 P = 90.23 R = **94.83**	A = 91.88 P = 91.69 R = 93.18	A = 92.46 P = 91.52 R = 94.63	A = 92.39 P = 91.53 R = 94.56	A = 92.57 P = 91.74 R = 94.71	A = **92.28** P = **91.90** R = 93.91	A = 92.06 P = 91.19 R = 94.49
Kitchen	A = 91.73 P = 90.23 R = **94.83**	A = 91.44 P = 91.08 R = 92.79	A = 92.07 P = **91.80** R = 93.22	A = **92.29** P = 91.34 R = 94.23	A = 91.95 P = 90.77 R = 94.66	A = 91.44 P = 90.77 R = 93.56	A = 91.73 P = 90.23 R = 94.83

We also ran this algorithm on a dataset with different product types: books, DVDs, electronics, and kitchen. Topic distribution features improved results here also.

6.4. Sentiment Analysis

6.4.1. Overall vs. Aspect-specific Sentiment Analysis

We compared overall sentiment prediction with an aspect specific one. To predict a sentiment for a specific aspect is more accurate rather than generating an overall prediction, regardless of the fact that it may be formed from a mixture of opinions on different aspects. Different aspects in the same review can contain different sentiment rankings. Some of them can be positive while others are negative.

Aspect sentiment can be described by a different lexicon then other aspects. For example, *"A cheap movie"* describes a negative sentiment about a quality of the movie, and *"I got cheap tickets"* describes a positive sentiment about a ticket price. In this example, a word *"cheap"* adds an ambiguity in overall sentiment prediction, while in aspect-specific sentiment prediction it plays a correct role.

We ran an SVM classification algorithm implemented by SVM-light with default parameters. Results of the comparison between overall and aspect-specific sentiment analysis are presented below:

Hotel Dataset

Table 27: Comparison between overall and aspect-specific sentiment analysis on the Hotel Dataset.

Aspect	*Best sentiment analysis results*
Overall	A = 76.061 , P= 74.624 , R= 80.091 (no topics)
Service	A = 81.251 , P= 78.379 , R= 86.462 (best 10 topics)
BService	A = 77.142 , P= 75.407 , R= 81.907 (best 60 topics)
Checkin	A = 85.318 , P= 83.651 , R= 88.297 (best 60 topics)
Value	A = 84.226 , P= 81.563 , R= 88.66 (best 10 topics)
Rooms	A = 82.019 , P= 78.977 , R= 87.447 (best 10 topics)
Clean	A = 81.664 , P= 78.748 , R= 87.059 (best 60 topics)
Location	A = 80.277 , P= 80.043 , R= 81.111 (best 100 topics)

In this dataset, the accuracy of an overall sentiment analysis is 76%, while an accuracy of aspect-specific sentiment analysis is between 77% and 85%. Precision and recall results of aspect-specific sentiment analysis are much higher than overall results. We can see that when testing for overall sentiment, topics did not improve the results. But, when looking at aspect-specific sentiment analysis, topic features consistently improve accuracy and F1-measure of sentiment analysis.

DVD Dataset

Table 28: Comparison between overall and aspect-specific sentiment analysis on the DVD Dataset.

Aspect	*Best sentiment analysis results*
Overall	A = 76.744 , P= 78.69 , R= 94.897 (no topics)
Audio	A = 84.838 , P= 84.931 , R= 99.616 (100 topics)
Extras	A = 70.518 , P= 70.509 , R= 83.335 (10 topics)
Movie	A = 72.504 , P= 72.549 , R= 99.349 (100 topics)
Video	A = 85.085 , P= 84.901 , R= 100.0 (100 topics)

In this dataset, sentiment analysis accuracy of *extras* and *movie* is lower than overall sentiment results. Probably, it is influenced by the small size of aspect-specific datasets, while an overall sentiment model is trained on a larger dataset. Also, in this dataset we can see that the best overall results were with no topics, topic distributions do not help in overall sentiment analysis.

Restaurants Dataset

Table 29: Comparison between overall and aspect-specific sentiment analysis on the Restaurant Dataset

Aspect	*Best sentiment analysis results*
Overall	A = 94.56 , P= 93.73 , R= 95.52 , F1 = 0.94
Food	A = 92.94 , P= 94.89 , R= 90.78 , F1 = 0.92
Ambience	A = 89.13 , P= 90.47 , R= 87.51 , F1 = 0.88

In this dataset, the overall aspect has higher results than food and ambience aspects, probably because this dataset is highly unbalanced.

Multi-Domain Dataset

Table 30: Comparison between overall and aspect-specific sentiment analysis on the Multi-Domain Dataset

Aspect	*Best sentiment analysis results*
Overall	A = 82.623 , P= 83.378 , R= 83.716
Books	A = 80.833 , P= 81.886 , R= 79.721
DVD	A = 72.916 , P= 77.722 , R= 65.0
Electronics	A = 83.25 , P= 85.242 , R= 81.0
Kitchen	A = 77.502 , P= 80.564 , R= 74.445

Because of the small size of this dataset, the overall aspect got higher results.

6.4.2. Aspect-specific sentiment analysis with topic distribution features

Here we describe the influence of topic distribution on aspect-specific sentiment analysis.

6.4.2.1. SVM Classification

We used an SVM binary classifier where the output of a learned function is either positive or negative. We ran the SVM-light classification package with default arguments.

Evaluation method

For each aspect, we performed two-stage algorithms. First, we performed a binary SVM aspect classification of a relevant aspect. All predicted review sentences were gathered for next sentiment classification. Second stage is an aspect sentiment classification performed on predicted review sentences using the SVM-light classification package.

Table 31: Sentiment classification results on the Hotels dataset.

Aspect	*No topics*	*10topics*	*60 topics*	*100 topics*
Service	A = 81.05 P = 79.07 R = 84.73 F1 = 81.80	A = **81.25** P = 78.37 R = 86.46 F1 = **82.20**	A = 80.88 P = 77.86 R = 86.45 F1 = 81.93	A = 80.31 P = 78.40 R = 84.27 F1 = 81.20
BService	A = 74.41 P = 72.22	A = 75.35 P = 73.98	A = **77.14** P = 75.40	A = 74.88 P = 73.04

	R = 80.00 F1 = 75.90	R = 78.57 F1 = 76.20	R = 81.90 F1 = **78.50**	R = 79.76 F1 = 76.20
Checkin	A = 83.36 P = 81.26 R = 87.11 F1 = 84.00	A = 83.93 P = 81.94 R = 87.44 F1 = 84.60	A = **85.31** P = 83.65 R = 88.29 F1 = **85.90**	A = 84.36 P = 82.66 R = 87.44 F1 = 84.90
Value	A = 83.61 P = 81.99 R = 86.42 F1 = 84.10	A = **84.22** P = 81.56 R = 88.66 F1 = **84.90**	A = 83.29 P = 82.18 R = 85.36 F1 = 83.70	A = 83.81 P = 81.29 R = 88.04 F1 = 84.50
Rooms	A = 80.26 P = 78.72 R = 83.19 F1 = 80.80	A = **82.02** P = 78.98 R = 87.45 F1 = **82.90**	A = 79.99 P = 78.05 R = 83.72 F1 = 80.70	A = 80.26 P = 78.67 R = 83.29 F1 = 80.90
Clean	A = 80.48 P = 78.92 R = 83.92 F1 = 81.30	A = 81.07 P = 78.43 R = 86.27 F1 = 82.10	A = **81.66** P = 78.74 R = 87.06 F1 = **82.60**	A = 81.37 P = 77.51 R = 88.62 F1 = 82.60
Location	A = 79.19 P = 78.26 R = 81.62 F1 = 79.90	A = 79.85 P = 80.06 R = 80.27 F1 = 80.10	A = 78.75 P = 78.50 R = 79.72 F1 = 79.10	A = **80.28** P = 80.04 R = 81.11 F1 = **80.50**

In this dataset, sentiment classification results of all aspects were improved using topics distribution features. For example, an accuracy of *BService* sentiment classification increased from 74.4% to 77.14%, and F1 score increased from 0.759 to 0.785. The accuracy of *Checkin* sentiment classification increased from 83.366% to 85.318%, and F1-score increased from 0.84 to 0.859.

Table 32: Sentiment classification results on Restaurants dataset.

Aspect	*No topics*	*10 topics*	*14 topics*	*20 topics*	*100 topics*
Food	A = **87.40** P = 88.41 R = 86.14 F1 = **87.20**	A = 86.63 P = 88.38 R = 84.37 F1 = 86.30	A = 85.66 P = 86.49 R = 84.59 F1 = 85.50	A = 86.74 P = 87.23 R = 86.15 F1 = 86.60	A = 87.18 P = 88.27 R = 85.85 F1 = 87.00
Ambience	A = 87.45 P = 88.57 R = 86.02 F1 = 87.20	A = 87.09 P = 88.02 R = 85.88 F1 = 86.90	A = **87.80** P = 89.13 R = 86.13 F1 = **87.60**	A = 87.41 P = 88.54 R = 85.99 F1 = 87.20	A = 87.35 P = 88.28 R = 86.15 F1 = 87.20

In this dataset, *Ambience* sentiment accuracy with 14 topic distribution was 87.8% compared to 87.45% accuracy without any topic distribution features. But, use of topic distribution features in *Food* sentiment did not improve sentiment classification accuracy.

Table 33: Sentiment classification results on DVD dataset.

Aspect	*no topics*	*10 topics*	*20 topics*	*100 topics*
Audio	A = 84.51 P = 84.53 R = 95.80 F1 = 90.50	A = 84.35 P = 84.28 R = 96.00 F1 = 90.47	A = 84.35 P = 84.39 R = 96.80 F1 = 90.40	A = **84.83** P = **84.93** R = **97.61** F1 = **91.60**
Extras	A = 70.17 P = 71.03 R = 80.91 F1 = 75.60	A = **70.52** P = **70.51** R = **83.34** F1 = **76.30**	A = 68.62 P = 69.60 R = 80.60 F1 = 74.00	A = 68.79 P = 69.26 R = 81.82 F1 = 75.00
Movie	A = 72.03 P = 72.20 R = 89.34 F1 = 80.60	A = 72.19 P = 72.10 R = 92.00 F1 = 81.00	A = 72.19 P = 72.10 R = 92.00 F1 = 81.70	A = **72.51** P = **72.55** R = **95.34** F1 = **83.80**
Video	A = 84.59	A = 84.26	A = 84.26	A = **85.09**

	P = 84.46 R = 95.03 F1 = 89.50	P = 84.18 R = 95.00 F1 = 89.40	P = 84.17 R = 96.00 F1 = 90.40	P = **84.90** R = **96.20** F1 = **90.80**

In this dataset, topic distribution improves aspect sentiment results in all aspects.

Table 34: Sentiment classification results on Multi-Domain dataset.

Product type	*No topics*	*10 topics*	*14 topics*	*20 topics*	*30 topics*	*50 topics*	*100 topics*
Books	A = 68.75 P = 72.80 R = 63.33 F1 = 67.70	A = 77.63 P = 79.93 R = 74.72 F1 = 77.20	A = 79.30 P = 80.33 R = 78.05 F1 = 79.10	A = 79.03 P = 80.47 R = 77.22 F1 = 78.80	A = **80.83** P = 81.88 R = 79.72 F1 = 78.80	A = 78.61 P = 80.49 R = 75.83 F1 = 78.0	A = 79.58 P = 81.37 R = 77.50 F1 = **79.30**
DVD	A = 68.75 P = 72.80 R = 63.33 F1 = 67.70	A = 71.25 P = 76.25 R = 63.33 F1 = 69.10	A = 71.24 P = 78.49 R = 60.00 F1 = 68.00	A = **72.92** P = 77.72 R = 65.00 F1 = **76.10**	A = 67.91 P = 71.19 R = 63.33 F1 = 67.00	A = 70.83 P = 76.67 R = 62.50 F1 = 68.80	A = 70.00 P = 73.84 R = 64.16 F1 = 68.60
Electronics	A = 81.12 P = 83.06 R = 79.00 F1 = 80.90	A = 82.87 P = 84.71 R = 81.25 F1 = 82.90	A = 81.62 P = 83.27 R = 80.5 F1 = 81.80	A = 82.00 P = 82.92 R = 81.50 F1 = 82.20	A = **83.25** P = 84.63 R = 81.75 F1 = **83.10**	A = 83.25 P = 85.24 R = 81.00 F1 = 83.00	A = 82.12 P = 82.92 R = 82.00 F1 = 82.40
Kitchen	A = 73.05 P = 75.19 R = 70.00 F1 = 72.50	A = **77.51** P = 80.56 R = 74.44 F1 = **77.30**	A = 77.5 P = 81.99 R = 71.11 F1 = 76.10	A = 75.55 P = 78.88 R = 71.11 F1 = 74.70	A = 76.94 P = 80.53 R = 72.22 F1 = 76.10	A = 75.55 P = 78.03 R = 72.22 F1 = 75.00	A = 75.83 P = 78.36 R = 72.22 F1 =0.75

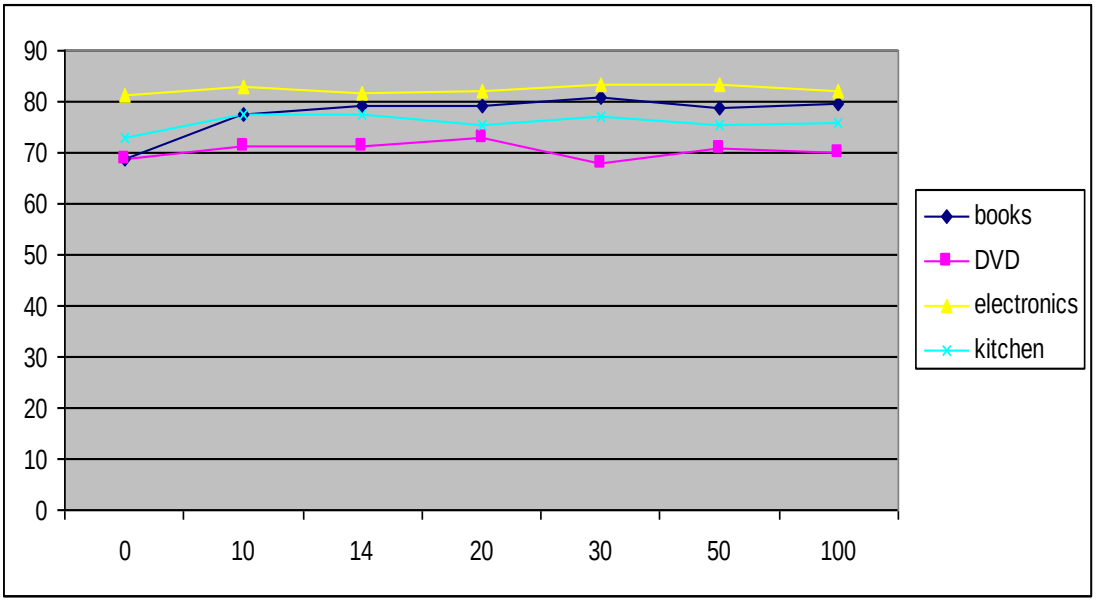

Figure 10: Graph of product sentiment classification accuracy.

In a multi-domain dataset, topic distribution features played a significant role in product-specific sentiment classification. In *Books*, use of 30 topic distribution features raised accuracy from 68.75% to 80.833%, and F1-score rose from 0.677 to 0.788.

In summary, topic distribution features improved base line in all reviewed datasets in almost all aspects/products.

6.4.2.2. SVM Regression

In order to compare with other methods, we also ran SVM regression to calculate mean squared difference (L2) and mean absolute difference (L1) between our method's prediction and true sentiment rating scores across test documents and aspects.

Evaluation Method

For each aspect, we performed two-stage algorithms. First, we performed binary SVM aspect classification of a relevant aspect. All predicted review sentences were gathered together for next sentiment classification. The second stage is an aspect-sentiment classification performed on predicted review sentences using the libSVM [Chang and Lin, 2011] regression package.

SVM Regression (SVR) is a method to estimate a function that maps from an input object to a real number based on training data. We use libSVM regression implementation with linear regression function and C parameter equal to 1 (these parameters showed best results).

We only compute these metrics for the DVD and Hotels dataset where the L1 and L2 measures have been reported in previous work. For other datasets, these metrics were not reported.

DVD Dataset

Table 35: SVM regression results on DVD dataset.

Aspect	*No topics*	*10 topics*	*20 topics*	*50 topics*	*100 topics*	*200 topics*
Audio	L1 = 0.892 L2 = 1.43	L1 = 0.878 L2 = 1.386	L1 = 0.889 L2 = 1.405	L1 = 0.892 L2 = 1.412	L1 = **0.879** L2 = **1.385**	L1 = 0.884 L2 = 1.420
Extras	L1 = 1.787 L2 = 5.097	L1 = 1.782 L2 = 5.067	L1 = 1.808 L2 = 5.163	L1 = 1.774 L2 = 5.158	L1 = 1.804 L2 = 5.173	L1 = **1.572** L2 = **3.986**
Movie	L1 = 1.687 L2 = 4.652	L1 = 1.691 L2 = 4.703	L1 = 1.702 L2 = 4.654	L1 = 1.697 L2 = 4.729	L1 = 1.692 L2 = 4.659	L1 = **1.648** L2 = **4.432**
Video	L1 = 1.026 L2 = 1.925	L1 = 1.026 L2 = 1.887	L1 = **1.016** L2 = **1.884**	L1 = 1.032 L2 = 1.895	L1 = 1.022 L2 = 1.890	L1 = 1.039 L2 = 1.912

Average over all aspects: L1 = 1.27, L2 = 2.92. (Results highlighted in red are best results for the aspect.)

In the DVD dataset, topic distribution features improve regression errors. For example, in the Extras aspect, regression mean absolute error without any topics distribution is 1.787, and with 200 topic distribution features is 1.572, while quadratic error improved from 5.097 to 3.986.

Hotels Dataset

Table 36: SVM regression results on Hotels dataset.

Aspect	*No topics*	*20 topics*	*30 topics*	*40 topics*	*60 topics*	*100 topics*
Service	L1 = 0.634 L2 = 0.869	L1 = 0.630 L2 = 0.846	L1 = **0.608** L2 = **0.808**	L1 = 0.620 L2 = 0.833	L1 = 0.628 L2 = 0.875	L1 = 0.625 L2 = 0.831
BService	L1 = 0.809 L2 = 1.207	L1 = 0.790 L2 = 1.176	L1 = **0.771** L2 = **1.122**	L1 = 0.774 L2 = 1.138	L1 = 0.809 L2 = 1.188	L1 = 0.792 L2 = 1.131
Checkin	L1 = 0.638	L1 = 0.642	L1 = 0.634	L1 = 0.645	L1 = 0.655	L1 = **0.619**

	L2 = 0.893	L2 = 0.903	L2 = 0.870	L2 = 0.888	L2 = 0.937	L2 = **0.785**
Value	L1 = 0.694 L2 = **0.849**	L1 = 0.694 L2 = 0.873	L1 = 0.686 L2 = 0.836	L1 = 0.686 L2 = 0.883	L1 = 0.691 L2 = 0.867	L1 = **0.683** L2 = 0.867
Rooms	L1 = 0.640 L2 = 0.741	L1 = 0.649 L2 = 0.750	L1 = **0.638** L2 = **0.725**	L1 = 0.651 L2 = 0.759	L1 = 0.643 L2 = 0.751	L1 = 0.641 L2 = 0.747
Clean	L1 = 0.468 L2 = 0.588	L1 = 0.465 L2 = 0.577	L1 = 0.461 L2 = 0.570	L1 = 0.461 L2 = **0.556**	L1 = **0.460** L2 = 0.567	L1 = 0.468 L2 = 0.572
Location	L1 = 0.514 L2 = 0.747	L1 = 0.515 L2 = 0.746	L1 = 0.514 L2 = 0.735	L1 = 0.520 L2 = 0.750	L1 = 0.513 L2 = 0.723	L1 = **0.507** L2 = **0.714**

Average over all aspects: L1 = 0.612, L2 = 0.798.

In this dataset, topic distribution features also improve regression errors.

6.5. Comparison with other Methods

6.5.1. Aspects Extraction

6.5.1.1. Restaurants Dataset

6.5.1.1.1. Comparison with ME-LDA

We compare our results with ME-LDA [Zhao et al., 2010] as the best result published on this dataset. ME-LDA is a joint Aspect and Sentiment unsupervised model.

Table 37: Comparison between ME-LDA and our method.

Aspect	*Method*	*Precision*	*Recall*	*F1*
Food	ME-LDA	0.874	0.787	0.828
	our approach	0.944	0.956	0.95
Ambience	ME-LDA	0.773	0.558	0.648
	our approach	0.945	0.956	0.950

Our method significantly outperformed the ME-LDA method.

6.5.1.1.2. Comparison with Brody and Elhadad, 2010

We compared our results with results reported in "An Unsupervised Aspect-Sentiment Model for Online Reviews". To determine the quality of their automatically inferred aspects for each sentence in the data, the LDA model assigns a distribution of topics over the set of inferred aspects. By defining a threshold for each aspect, they labeled a sentence as belonging to a specific aspect if an assigned distribution is bigger than the threshold. By varying the threshold they created precision-recall curves for the top three ratable aspects in the restaurant domain:
In the Food aspect, when precision is greater than 90%, recall is less than 70%. In our approach, both precision and recall are higher than 90%. In Atmosphere and Service aspects, the results are similar.

6.5.2. Aspects-specific Sentiment Classification

6.5.2.1. ME-LDA model

The ME-LDA model separates aspect and opinion words. They quantitatively evaluated the quality of the aspect specific opinion words identified by ME-LDA. They do not evaluate aspect-specific sentiment of entire reviews.

6.5.2.2. Comparison with [Brody and Elhadad, 2010]

We compared our results with results reported [Brody and Elhadad, 2010]. Brody and Elhadad did not combine the sentiment and aspect components into a full-fledged aspect-sentiment review analysis system, and did not conduct an aspect-specific sentiment analysis evaluation of entire reviews.

6.5.2.3. DVD dataset

We compared with [Sauper et al., 2010]. In the Joint Content Model, the authors obtain a baseline system by eliminating content features and only using a task model with the set of features described above. They also compare against a simplified variant of their method wherein a content model is induced in isolation rather than learned jointly in the context of the underlying task. In their experiments, they refer to the two methods as the No Content Model (NoCM) and Independent Content Model (IndepCM) settings, respectively. The Joint Content Model (JointCM) setting refers to their full model described in Section 4.7, where content and task components are learned jointly.

Table 38: Evaluation results on the multi-aspect sentiment ranking of JointCM approach.

	L_1	L_2
NoCM	1.37	3.15
IndepCM	1.28†*	2.80†*
JointCM	**1.25**†	**2.65**†*
Gold	1.18†*	2.48†*

Our approach reported L1 = 1.27, L2 = 2.92. L1 is better than IndepCM but slightly worse than JointCM; L2 is better than NoCM but worse than others. The quality of our method is determined by the quantity of training data. Each aspect has only 665 review sentences, while number of Audio negative reviews is 103 and positive is 527.

6.5.2.4. Hotels Dataset

We compared with the method of "Multi-facet Rating of Product Reviews" [Baccianella et al., 2009]. Their best L1 result for average results across the seven aspect-specific datasets is 0.733 while we reported 0.612.

Chapter 7:Conclusions

7.1. Summary of Method

We propose a two-step approach for aspect-sentiment classification. Following linguistic intuition and previous research, rich contextual information can benefit text analysis application such as aspect extraction and sentiment classification. The induced content structure is learned from a large un-annotated corpus using unsupervised topic modeling.

- Aspect classification
  In order to explore aspects in the reviews, we used a supervised learning method with unsupervised features. We use a Support Vector Machine, where vector features are constructed from unigrams, bi-grams, and topic distributions are inferred by a local version of LDA for each sentence in the data.
- Aspects Sentiment extraction
  We used the same features as in aspect extraction: unigrams, bigrams, and topics distribution over each sentence. Topics are useful latent structures to explain semantic association. Latent topics are discovered by identifying groups of words in the corpus that frequently occur together within documents.

7.2. Contributions

We tested our method on four different data sets. We used the same metrics in order to compare and investigate the influence of different data sources on our method's results. We introduced a new 2-stage method of sentiment classification using topic distributions as an SVM feature.

7.3. Results

Aspect extraction using our method showed better results than base line (just unigrams) and than ME-LDA [Titov and McDonald, 2008] and LocLDA [Brody and Elhadad, 2010]. For aspect-sentiment classification, our method improved our base line and showed better results than [Baccianella et al., 2009] on the same Hotels dataset. On the DVD dataset, we reported better results than our baseline, but [Sauper et al., 2010] reported better results than us on their joint model.

7.4. Discussion

In our work, we investigated how modeling content structure can benefit test analysis applications such as sentiment analysis and how aspect specific sentiment analysis can be more accurate than global sentiment analysis.

One of the future directions we plan to explore is joint models that jointly learn aspects and sentiments. Another direction is to use this method on different text analysis applications as extractive summarization.

Chapter 8: Bibliography

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. Multi-facet Rating of Product Reviews. *ECIR '09 Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, 2009.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research 3:993–1022,* 2003.

John Blitzer, Mark Dredze, Fernando Pereira. Domain Adaptation for Sentiment Classification. *Association of Computational Linguistics* (ACL), 2007

Samuel Brody, Noemie Elhadad. An Unsupervised Aspect-Sentiment Model for Online Reviews. *The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *Transactions on Intelligent Systems and Technology*, 2011.

Kushal Dave, Steve Lawrence, David Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *Proceedings of the 12th international conference on World Wide Web*, 2003

Adnan Duric and Fei Song. Feature Selection for Sentiment Analysis Based on Content and Syntax Models. *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis,* 2011.

Brian Eriksson. Sentiment classification of movie reviews using linguistic parsing.

Gayatree Ganu, Noémie Elhadad, and Amélie Marian. Beyond the Stars: Improving Rating Predictions using Review Text Content. *WebDB. Providence, RI*, 2009.

Stephan Green, Philip Resnik. Syntactic Packaging and Implicit Sentiment. *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009.

Vasileios Hatzivassiloglou, Kathleen R. McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, 1997.

Hu, Minqing and Bing Liu. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2004.

Kim, Soo-Min and Eduard Hovy. Determining the Sentiment of Opinions. *Proceedings of COLING-04. pp. 1367-1373. Geneva, Switzerland*, 2004

Nathalie Japkowicz and Shaju Stephen. The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis , Volume 6 Issue 5*, 2002.

Valentin Jijkoun and Katja Hofmann. Generating a Non-English Subjectivity Lexicon: Relations That Matter. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009.

Yohan Jo, Alice Oh. Aspect and Sentiment Unification Model for Online Review Analysis. *WSDM '11: Proceedings of the fourth ACM international conference on Web search and data mining*, 2011

Thorsten Joachims. SVM-light implementation. 2008

Chenghua Lin, Yulan He. Joint Sentiment/Topic Model for Sentiment Analysis. *CIKM **'09:** Proceeding of the 18th ACM conference on Information and knowledge management*, 2009.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning Word Vectors for Sentiment Analysis. *HLT '11: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, 2011.

Shotaro Matsumoto, Hiroya Takamura, Manabu Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. *PAKDD'05: Proceedings of the 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining*, 2005.

McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit."

Miller, G., R, Beckwith, C. Fellbaum, D. Gross, and K. Miller. Wordnet: an on-line lexical database. *Oxford Univ. Press.*, 1990.

Mohammad, S., C. Dunne., and B. Dorr. Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*, 2009.

Katharina Morik and Peter Brockhausen and Thorsten Joachims. Combining statistical learning with a knowledge-based approach: A case study in intensive care monitoring. *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, 1999.

Bo Pang, Lillian Lee, Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, 2002.

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

Popescu, Ana-Maria and Oren Etzioni. 2005. Extracting product features and opinions from reviews. *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005.

Christina Sauper, Aria Haghighi, Regina Barzilay. Incorporating Content Structure into Text Analysis Applications. *EMNLP '10: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

Mostafa Al Masum Shaikh, Helmut Prendinger, Ishizuka Mitsuru. Assessing Sentiment of Text by Semantic Dependency and Contextual Valence Analysis. *ACII '07: Proceedings of the 2nd international conference on Affective Computing and Intelligent Interaction*, 2007.

S. Shivashankar, B. Ravindran. Multi Grain Sentiment Analysis using Collective Classification. *Proceedings of the 2010 conference on ECAI 2010: 19th European Conference on Artificial Intelligence*, 2010.

Ivan Titov, Ryan McDonald. Modeling Online Reviews with Multi-grain Topic Models. *WWW '08: Proceeding of the 17th international conference on World Wide Web*, 2008

Theresa Wilson, Janyce Wiede, Paul Hoffman. Recognizing contextual polarity in Phrase-Level Sentiment Analysis. *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005

Hongning Wang, Yue Lu, Chengxiang Zha. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010

Yong Zhang, Dong-Hong Ji, Ying Su and Po Hu. Topic Analysis for Online Reviews with an Author-Experience-Object-Topic Model. *AIRS'11: Proceedings of the 7th Asia conference on Information Retrieval Technology*, 2011.

Zhongwu Zhai, Bing Liu, Hua Xu, Peifa Jia. Constrained LDA for Grouping Product Features in Opinion Mining. *PAKDD'11: Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining - Volume Part I*, 2011.

Tian-Jei Zhan and Chun-Hung Li. Semantic Dependent Word Pairs Generative Model for Fine-Grained Product Feature Minung. *PAKDD'11 Proceedings of the 15th Pacific-Asia conference on Advances in knowledge discovery and data mining - Volume Part I*, 2011.

Wayne Xin Zhao, Jing Jiang, Hongfei Yan, Xiaoming Li. Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid. *EMNLP '10: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.