DENIS DERATANI MAUÁ

MODELOS DE TÓPICOS NA CLASSIFICAÇÃO AUTOMÁTICA DE RESENHAS DE USUÁRIO

São Paulo
2009

DENIS DERATANI MAUÁ

MODELOS DE TÓPICOS NA CLASSIFICAÇÃO AUTOMÁTICA DE RESENHAS DE USUÁRIO

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção do título de Mestre em Engenharia

São Paulo
2009

DENIS DERATANI MAUÁ

MODELOS DE TÓPICOS NA CLASSIFICAÇÃO AUTOMÁTICA DE RESENHAS DE USUÁRIO

Dissertação apresentada à Escola Politécnica da Universidade de São Paulo para a obtenção do título de Mestre em Engenharia

Área de concentração:
Engenharia de Controle e Automação Mecânica

Orientador:
Prof. Dr. Fabio Gagliardi Cozman

São Paulo
2009

AGRADECIMENTOS

Ao Professor Fabio Cozman, pela paciência e dedicação na orientação deste trabalho, por seus sábios conselhos e valorosas sugestões no texto, por estar aberto às discussões e às idéias, pela acessibilidade e, principalmente, por ter-me proporcionado esta oportunidade.

A meus colegas de laboratório, Rodrigo, Victor, José, Ricardo e Daniel, pelas muitas ajudas que me prestaram, pelas discussões, dicas e, acima de tudo, pela amizade.

A meus grandes amigos Wagner e Ivo, que durante todo este tempo de Mestrado, mesmo quando me dizendo o contrário, me motivaram a continuar. Por suas agradáveis companhias na reposição da cafeína tão necessária para o trabalho de pesquisa, por partilharem suas opiniões e me deixarem partilhar as minhas. À minha amiga Kelly, que, ao lado de outras pessoas queridas que conheci aqui, fez minha estadia nesta cidade inóspita mais amena.

A meus pais, por me proverem o essencial sem o qual nada disto teria sido possível.

À Karina, pela compreensão, motivação, inteligência e bom-humor dedicados a mim. Por me fazer parte da vida dela e preencher a minha vida. Pela companhia sem a qual todo o tempo perdido neste trabalho e em qualquer outro não teria e não terá verdadeiramente valor.

RESUMO

Existe um grande número de resenhas de usuário na internet contendo valiosas informações sobre serviços, produtos, política e tendências. A compreensão automática dessas opiniões é não somente cientificamente interessante, mas potencialmente lucrativa.

A tarefa de classificação de sentimentos visa a extração automática das opiniões expressas em documentos de texto. Diferentemente da tarefa mais tradicional de categorização de textos, na qual documentos são classificados em assuntos como esportes, economia e turismo, a classificação de sentimentos consiste em anotar documentos com os sentimentos expressos no texto. Se comparados aos classificadores tradicionais, os classificadores de sentimentos possuem um desempenho insatisfatório. Uma das possíveis causas do baixo desempenho é a ausência de representações adequadas que permitam a discriminação das opiniões expressas de uma forma concisa e própria para o processamento de máquina.

Modelos de tópicos são modelos estatísticos que buscam extrair informações semânticas ocultas na grande quantidade de dados presente em coleções de texto. Eles representam um documento como uma mistura de tópicos, onde cada tópico é uma distribuição de probabilidades sobre palavras. Cada distribuição representa um conceito semântico implícito nos dados. Modelos de tópicos, as palavras são substituídas por tópicos que representam seu significado de forma sucinta. De fato, os modelos de tópicos realizam uma redução de dimensionalidade nos dados que pode levar a um aumento do desempenho das técnicas de categorização de texto e recuperação de informação. Na classificação de sentimentos, eles podem fornecer a representação necessária através da extração de tópicos que representem os sentimentos expressos no texto.

Este trabalho dedica-se ao estudo da aplicação de modelos de tópicos na representação e classificação de sentimentos de resenhas de usuário. Em particular, o modelo Latent Dirichlet Allocation (LDA) e quatro extensões (duas delas desenvolvidas pelo autor) são avaliados na tarefa de classificação de sentimentos baseada em múltiplos aspectos. As extensões ao modelo LDA permitem uma investigação dos efeitos da incorporação de informações adicionais como contexto, avaliações de aspecto e avaliações de múltiplos aspectos no modelo original.

Palavras-chave: inteligência artificial, aprendizado computacional, aprendizado de máquina, processamento de texto, categorização de texto, classificação de sentimento.

ABSTRACT

There is a large number of user reviews on the internet with valuable information on services, products, politics and trends. There is both scientific and economic interest in the automatic understanding of such data.

Sentiment classification is concerned with automatic extraction of opinions expressed in user reviews. Unlike standard text categorization tasks that deal with the classification of documents into subjects such as sports, economics and tourism, sentiment classification attempts to tag documents with respect to the feelings they express. Compared to the accuracy of standard methods, sentiment classifiers have shown poor performance. One possible cause of such a poor performance is the lack of adequate representations that lead to opinion discrimination in a concise and machine-readable form.

Topic Models are statistical models concerned with the extraction of semantic information hidden in the large number of data available in text collections. They represent a document as a mixture of topics, probability distributions over words that represent a semantic concept. According to Topic Model representation, words can be substituted by topics able to represent concisely its meaning. Indeed, Topic Models perform a data dimensionality reduction that can improve the performance of text classification and information retrieval techniques. In sentiment classification, they can provide the necessary representation by extracting topics that represent the general feelings expressed in text.

This work presents a study of the use of Topic Models for representing and classifying user reviews with respect to their feelings. In particular, the Latent Dirichlet Allocation (LDA) model and four extensions (two of them developed by the author) are evaluated on the task of aspect-based sentiment classification. The extensions to the LDA model enables us to investigate the effects of the incorporation of additional information such as context, aspect rating and multiple aspect rating into the original model.

Keywords: artificial intelligence, computational learning, machine learning, text processing, text categorization, sentiment classification.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

T - Número de tópicos, p. 33

φ_j - Distribuição sobre um vocabulário fixo relativa ao j-ésimo tópico, p. 34

β - Hiperparâmetro de suavização dos tópicos, p. 34

θ_d - Distribuição de tópicos relativo ao d-ésimo documento, p. 34

α - Hiperparâmetro de suavização das distribuições de tópicos, p. 34

z_i - Rótulo de tópico da i-ésima palavra na coleção, p. 34

w_i - i-ésima palavra da coleção, p. 34

W - Número de palavras distintas na coleção, p. 34

$n_j^{(k)}$ - Número de palavras iguais ao k-ésimo vocábulo e atribuídas ao j-ésimo tópico na coleção, p. 34

D - Total de documentos na coleção, p. 34

$n_j^{(d)}$ - Número de palavras atribuídas ao j-ésimo tópico no d-ésimo documento, p. 34

N - Número de palavras na coleção, p. 35

$n_{-i,j}^{(k)}$ - Número de palavras iguais ao k-ésimo vocábulo e atribuídas ao j-ésimo tópico na coleção, desconsiderando a i-ésima palavra, p. 36

$n_{-i,j}^{(d)}$ - Número de palavras atribuídas ao j-ésimo tópico no d-ésimo documento, desconsiderando a i-ésima palavra, p. 36

N_d - Número de palavras no d-ésimo documento, p. 37

$n_j^{(k,k')}$ - Número de palavras iguais ao k-ésimo vocábulo precedidas por palavras iguais ao k'-ésimo vocábulo e assinaladas ao tópico j, p. 40

$\varphi_{(j,k)}$ - Distribuição sobre um vocabulário fixo relativa ao j-ésimo tópico e condicionadas no k-ésimo vocábulo, p. 40

η - Parâmetro de suavização das distribuições de rótulos, p. 41

n_ℓ - Número de documentos rotulados como ℓ, p. 43

A - Número de aspectos, p. 46

K - Número de notas, p. 46

$\varphi_{(j,y)}$ - Distribuição sobre um vocabulário fixo em relação ao aspecto j e nota y, p. 46

$r_j^{(d_i)}$ - Nota em relação ao aspecto j no documento ao qual pertence a i-ésima palavra, p. 46

a_i - Aspecto da i-ésima palavra, p. 46

$n_{-i,(j,y)}^{(k)}$ - Número de vezes que o k-ésimo vocábulo é assinalado ao aspecto j e nota y, não incluindo a i-ésima palavra, p. 47

$n_{-i,j}^{(k,d)}$ - Número de palavras k rotuladas atribuídas ao j-ésimo aspecto no d-ésimo documento, p. 48

SUMÁRIO

1 INTRODUÇÃO

Existe um imenso e valioso banco de dados de opiniões na internet, compreendendo de análises políticas a avaliações sobre produtos e serviços (PANG; LEE, 2008). Essa imensa quantidade de dados está, em grande parte, disponível somente na forma de textos escritos em linguagem humana, o que a torna inacessível para grande parte das aplicações de mineração de dados, que requerem dados em forma estruturada.

Boa parte dessas opiniões estão contidas em resenhas escritas por usuários. Em particular, sítios como *Yelp* (`http://www.yelp.com`), *We8there* (`http://www.we8there.com`), *Epinions* (`http://www.epinions.com`) e *Amazon* (`http://www.amazon.com`) incentivam a participação dos usuários através da postagem de resenhas sobre determinados produtos ou serviços. Em tais ambientes é comum deparar-se com centenas e até milhares de resenhas para um objeto ou classe de objetos, o que torna a coleta de opiniões por um interessado em um determinado item tediosa e facilmente incompleta.

De forma a superar essa sobrecarga de informação, pesquisadores da área de *análise de sentimentos* dedicam-se ao desenvolvimento de técnicas que permitam representar sucintamente os sentimentos expressos em um ou mais documentos (TURNEY, 2001; PANG; LEE; VAITHYANATHAN, 2002; PANG, 2005). A abordagem mais usualmente adotada, denominada *classificação de sentimentos*, consiste em classificar documentos de texto em categorias representativas de uma opinião (e.g., duas classes indicando se o texto exprime em geral um sentimento positivo ou negativo). Se comparada às tarefas tradicionais de categorização de texto, como detecção de *spams* em *emails* ou indexação automática de documentos por assunto (MANNING; SCHÜTZE, 1999; SEBASTIANI, 2002), essa abordagem apresenta um desempenho insatisfatório (PANG; LEE, 2008).

1.1 Resenhas de usuário

Uma resenha de usuário é um documento de texto contendo opiniões pessoais sobre um objeto. Os Exemplos 1 e 2 ilustram duas resenhas de usuário contendo opiniões

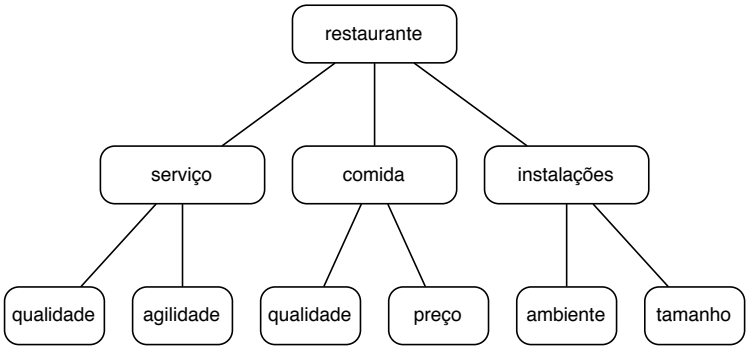


Figura 1: Exemplo de objeto com suas componentes e características ordenadas hierarquicamente. Cada nó não raíz do grafo denota um aspecto. Nós não terminais resumem o sub-conjunto de componentes e características dado pelos seus filhos.

antagônicas sobre um determinado restaurante.

Exemplo 1. *Embora a decoração seja ruim, a comida vale a pena.*

Exemplo 2. *Embora a comida seja ruim, a decoração vale a pena.*

Usa-se o termo *objeto* para denotar uma entidade alvo da avaliação de um usuário numa resenha, como por exemplo um produto (e.g., um carro ou um livro) ou um prestador de serviço (e.g., um restaurante ou um provedor de acesso à internet). Um objeto possui um conjunto de componentes (partes) e um conjunto de características (propriedades) (DING; LIU; YU, 2008). A Figura 1 ilustra um objeto do tipo restaurante e algumas de suas componentes e propriedades. Como mostra a figura, as partes e propriedades de um objeto podem ser agrupadas hierarquicamente, de forma que uma componente ou característica resuma um subconjunto de componentes ou características.

Definição 1. *Um* objeto *O é uma árvore cuja raíz representa o próprio objeto e cada nó não raíz representa uma componente ou característica do objeto.*[1]

Denota-se *aspecto* qualquer componente ou característica de um objeto que possa ser avaliada pelo usuário. Por exemplo, na Figura 1, *qualidade da comida*, *serviço* e *tamanho das instalações* são diferentes exemplos de aspectos de um objeto restaurante.

Definição 2. *Um* aspecto *A é um nó não raíz de um objeto O.*

As técnicas de classificação de sentimentos assumem, de maneira geral, que uma resenha pode ser concisamente representada por uma única métrica de avaliação, capaz

[1] Uma árvore é um grafo conectado e acíclico.

de exprimir o sentimento mais proeminente em um texto. Entretanto, como notado em Snyder e Barzilay (2007) e em Titov e McDonald (2008b), as opiniões expressas pelos autores são comumente multifacetadas, o que as impossibilita de serem representadas por uma única nota ou métrica. Tomem-se as resenhas dos Exemplos 1 e 2. Nelas, pode-se observar a ocorrência de sentimentos opostos em relação a diferentes aspectos de um objeto. Se se quisesse representar essas resenhas através de uma simples métrica, evitando classificações tendenciosas, dever-se-ia atribuir uma categoria que denotasse um sentimento neutro a elas, o que descartaria de fato as opiniões contidas nas resenhas.

Uma abordagem mais satisfatória à representação de textos opinativos é, portanto, a classificação de sentimentos em relação a diferentes aspectos de um objeto. Por exemplo, as resenhas dos Exemplos 1 e 2 poderiam ser classificadas quanto ao seu sentimento em relação aos aspectos *qualidade da comida* e *decoração*, como mostram os Exemplos 3 e 4, resultando numa representação concisa mais significativa dos dados originais. Esta abordagem é conhecida na literatura pelo nome de *classificação de sentimentos baseada em aspectos* (SNYDER; BARZILAY, 2007; TITOV; MCDONALD, 2008b).

Exemplo 3. decoração*: ruim,* qualidade da comida*: bom*

Exemplo 4. decoração*: bom,* qualidade da comida*: ruim*

Cabe aqui uma breve discussão sobre a terminologia adotada. O termo *palavra* pode ser utilizado para descrever pelo menos dois conceitos distintos: o de um tipo geral de coisa e o de uma instância particular (MANNING; SCHÜTZE, 1999). Neste texto adota-se esta última definição para palavra, a de unidade básica de um documento de texto, e reserva-se o termo *vocábulo* para descrever a identidade de uma palavra. Esta nomenclatura assemelha-se a comumente adota em processamento de texto que define uma instância particular como um *word token* (ocorrência de palavra) e uma identidade como um *word type* (tipo de palavra). O verso a seguir, retirado do poema *Sacred Emily* de Gertrude Stein, ilustra bem a diferença entre os conceitos.

> Rose is a rose is a rose is a rose.

No trecho existem, de acordo com a nomenclatura adotada, três vocábulos e dez palavras. A fim de se obter maior ênfase na distinção, por vezes utiliza-se o termo *ocorrência de uma palavra* ou simplesmente *ocorrência* com o mesmo sentindo do termo palavra. Vale notar que o termo vocábulo não é adotado aqui com o sentido de conceito semântico ou vocábulo linguístico. Assim, o vocábulo *manga* é utilizado como representação tanto do

conceito de fruta como do conceito de parte de uma veste. Por fim, o termo *vocabulário* é usado para descrever o conjunto de todos os vocábulos utilizados numa coleção de documentos de texto, ou seja, o conjunto de palavras distintas presentes na coleção.

1.2 Bag of words

Em sua forma "bruta", documentos de texto armazenados em um computador não são mais que listas de códigos alfanuméricos. Para que sejam utilizados por um algoritmo de aprendizado é necessário, antes de tudo, que estes dados "brutos" sejam convertidos numa representação adequada. Grande parte da pesquisa em classificação de texto e, mais especificamente, em classificação de sentimentos dedica-se ao estudo de novas e melhores formas de representação de textos (PANG; LEE; VAITHYANATHAN, 2002; PANG, 2005; SNYDER; BARZILAY, 2007; PANG; LEE, 2008; TITOV; MCDONALD, 2008a, 2008b).

Os métodos de classificação de texto baseados em técnicas de *aprendizado de máquina* comumente adotam um modelo de representação denominado *bag of words*, que consiste em representar documentos como vetores cujos elementos denotam o número de ocorrências de cada vocábulo no texto (SEBASTIANI, 2002). Este modelo implica no descarte de qualquer informação de dependência semântica entre as palavras. O Exemplo 5 ilustra as resenhas dos Exemplos 1 e 2 representadas pelo modelo *bag of words*. Ao descartar-se a ordem das palavras, ambos os documentos resultam numa mesma representação.

Exemplo 5. *a a a comida decoração embora pena ruim seja vale*

A hipótese por trás da representação *bag of words* é a de que documentos pertencentes a assuntos distintos possuam distribuições empíricas de vocábulos (significativamente) diferentes. Por exemplo, em artigos esportivos espera-se, segundo essa representação, encontrar um número maior de ocorrências do vocábulo "atleta" do que em artigos econômicos. Desta forma, é possível predizer o assunto principal ao qual se refere um documento através da simples verificação das freqüências de ocorrência. Apesar de simplista, essa hipótese mostrou-se bem sucedida na tarefa de classificação de textos em assuntos, exibindo desempenho superior ao de representações mais complexas (SEBASTIANI, 2002).

Entretanto, ao modelar-se documentos para a classificação de sentimentos, esta hipótese simplista pode degradar seriamente o desempenho, pois agora não se está interessado em determinar o assunto principal ao qual o texto pertence (e.g., restaurantes, hotéis, produtos eletrônicos), mas sim a inclinação das diversas "micro"-opiniões presentes (e.g.,

positiva em relação ao aspecto *comida*). Pode-se conceber toda a coleção de resenhas de usuários como pertencentes a um só tópico, o de resenhas de usuários, ou ainda, de textos avaliativos, esperando-se dessa forma encontrar distribuições semelhantes. Esse fato pode explicar a causa do baixo desempenho de classificadores de sentimento baseados no modelo *bag of words* (PANG; LEE, 2008).

Em boa parte, o problema introduzido pela representação *bag of words* pode ser solucionado anotando-se as palavras com os tópicos ou aspectos aos quais elas se referem. Os Exemplos 6 e 7 apresentam uma possível representação alternativa para as resenhas dos Exemplos 1 e 2, onde as palavras são rotuladas com aspectos. Os índices A, C e N indicam que a palavra refere-se, respectivamente, aos aspectos *ambiente*, *comida* ou *nenhum destes*. Note que nessa representação as resenhas são mapeadas em vetores distintas. A anotação das palavras com os aspectos referidos introduz parte das relações de dependência entre as palavras presentes no texto original, fazendo com que as resenhas sejam mapeadas em vetores distintos, diferentemente da representação *bag of words* do Exemplo 5.

Exemplo 6. $a^A\ a^C\ a^C\ comida^C\ decoração^A\ embora^N\ pena^C\ ruim^A\ seja^A\ vale^C$

Exemplo 7. $a^A\ a^A\ a^C\ comida^C\ decoração^A\ embora^N\ pena^A\ ruim^C\ seja^C\ vale^A$

Uma maneira de se realizar a anotação automaticamente é através da construção de um classificador capaz de determinar para cada trecho do texto (palavra, frase, sentença ou parágrafo) quais os aspectos opinados. No entanto, as técnicas tradicionais de construção de classificadores requerem o desenvolvimento de um conjunto de dados rotulados, cuja construção pode demandar um enorme tempo (MANNING; SCHÜTZE, 1999). De forma a evitar fenômenos indesejados que levem a um desempenho ruim, como por exemplo o "overfitting" do modelo aprendido,[2] recomenda-se utilizar uma quantidade de dados rotulados pelo menos duas vezes maior que a dimensão do espaço de atributos onde os dados são representados (SEBASTIANI, 2002; HASTIE; TIBSHIRANI; FRIEDMAN, 2001). Devido ao tamanho dos vocabulários empregados (usualmente de mais de 5000 vocábulos) problemas de processamento estatístico de linguagem requerem comumente um conjunto muito grande de dados rotulados (> 10000).

De forma a avaliar o grau de dificuldade de obtenção de um conjunto de dados rotulados para a tarefa de classificação de palavras em aspectos, o autor deste trabalho realizou

[2] O problema de "overfitting" ocorre quando o algoritmo de estimação não consegue generalizar suficientemente os parâmetros aprendidos a partir dos dados e incorpora o ruído aleatório no modelo final, levando a uma queda do desempenho na predição de novos dados.

uma etiquetação manual de um conjunto de resenhas de usuários extraídas da internet, com a ajuda de um aplicativo especialmente desenvolvido para a tarefa. A etiquetação de pouco mais de 200 resenhas e aproximadamente 1600 palavras tomou cerca de 30 dias de trabalho exclusivo, demonstrando a grande dificuldade desta abordagem.

Uma alternativa que dispensa a utilização de dados rotulados é a utilização de algoritmos de *aprendizado não supervisionado*, nos quais os parâmetros são estimados a partir dos dados utilizando um critério de otimalidade, evitando assim a necessidade de rotulação prévia (HASTIE; TIBSHIRANI; FRIEDMAN, 2001). Uma classe de modelos não supervisionados particularmente interessante para a tarefa de segmentação de texto em classes semânticas é a dos chamados *modelos de tópicos* (BLEI et al., 2003; GRIFFITHS; STEYVERS, 2004; STEYVERS; GRIFFITHS, 2007).

1.3 Modelos de tópicos

Modelos de tópicos são modelos probabilísticos generativos[3] que visam representar membros de uma coleção de documentos (de texto) através de descrições sucintas que mantenham as relações estatísticas suficientes para tarefas como classificação de textos, sumarização e recuperação de informação (BLEI et al., 2003; STEYVERS; GRIFFITHS, 2007). Como detalhado no Capítulo 3, os modelos de tópicos assumem que um documento pode ser adequadamente representado por um vetor cujos elementos indicam a proporção de palavras presentes no documento referentes a um tópico, uma distribuição de probabilidades sobre um vocabulário fixo que representa um conceito semântico ou assunto. Quando aplicados ao domínio de resenhas de usuários, esses tópicos podem corresponder aos diferentes aspectos comentados, aos diferentes sentimentos, ou ainda aos diferentes assuntos, especialidades ou locais citados no texto.

De forma a ilustrar a representação realizada pelos modelos de tópicos, assuma que na coleção formada pelas resenhas dos Exemplos 1 e 2 existem quatro tópicos referentes, respectivamente, a opiniões negativas e positivas em relação aos aspectos *comida* e *ambiente*. Os Exemplos 8 e 9 apresentam uma possível atribuição de palavras a tópicos, supostamente estimada a partir dos dados.

Exemplo 8. *Embora*^{3} *a*^{3} *decoração*^{3} *seja*^{3} *ruim*^{3}*,* *a*^{2} *comida*^{2} *vale*^{2} *a*^{2} *pena*^{2}*.*

Exemplo 9. *Embora*^{1} *a*^{1} *comida*^{1} *seja*^{1} *ruim*^{1}*,* *a*^{4} *decoração*^{4} *vale*^{4} *a*^{4} *pena*^{4}*.*

[3]Um modelo generativo descreve o processo de geração dos dados observados através da especificação de uma distribuição conjunta. Eles contrastam com os modelos discriminativos que modelam diretamente o processo de decisão especificando uma distribuição condicional nos dados observados.

Tabela 1: Amostra de tópicos estimados com o modelo Nota-Aspecto e 20 tópicos no conjunto de dados we8there. As dez palavras mais discriminativas são mostradas em ordem decrescente de pontuação para cada tópico.

Tópico	**Palavras melhor pontuadas**
comida-	ruim terrível insossa fria falaram boa como não pior dinheiro
comida+	deliciosa boa ótima pequeno apimentada usualmente fazer variedade grelhado pratos
serviço-	minutos fizeram garçonete levou pegou atendente finalmente perguntou questionário devagar
serviço+	ótimo excelente melhor simpático equipe familiar lugar ambiente molho espetacular

De acordo com a representação adotada nos modelos de tópicos, os documentos são mapeados em um vetor $\mathbf{v} = \langle n_1, n_2, n_3, n_4 \rangle$, onde a notação n_i denota o número de palavras assinaladas ao tópico i em um documento. Desta forma, considerando-se a rotulação dos Exemplos 8 e 9, as Resenhas 1 e 2 são representadas, respectivamente, no novo espaço definido pelo modelo de tópicos pelos pontos $\mathbf{r}_1 = \langle 0, 5, 5, 0 \rangle$ e $\mathbf{r}_2 = \langle 5, 0, 0, 5 \rangle$. Note que os documentos são mapeados em pontos distintos, de forma similar à representação dos Exemplo 6 e 7.

A idéia por trás da representação dos modelos de tópico é a de revelar a estrutura latente escondida nos dados, convertendo os dados brutos em uma representação de baixa dimensionalidade representativa de aspectos semânticos do texto. A representação é, desta forma, particularmente útil para remover fenômenos de ambigüidade léxica presentes no texto, como a polissemia (múltiplos significados para uma palavra) e a sinonímia (múltiplas palavras para um significado), que podem degradar o desempenho de classificadores de texto. Em especial na tarefa de classificação de sentimentos baseada em aspectos, ela proporciona uma solução para a ambigüidade introduzida pelo descarte da informação posicional e relacional (i.e., de dependência entre as palavras) presente no modelo *bag of words*.

A Tabela 1 reproduz, para fins ilustrativos, uma amostra de quatro dos vinte tópicos estimados a partir de um conjunto de dados reais de resenhas de usuários extraídas da internet (Capítulo 5) utilizando o modelo Nota-Aspecto (Capítulo 3). A lista completa de tópicos encontra-se na Tabela 22 do Anexo A. Na Tabela 1, os tópicos estão representados pelas dez palavras mais discriminativas. Os dois primeiros tópicos referem-se a avaliações negativas e positivas, respectivamente, relativas ao aspecto *comida*. Os dois últimos tópicos referem-se a diferentes avaliações em relação ao aspecto *serviço*. As palavras foram traduzidas do inglês para o português.

Os modelos de tópicos possibilitam não apenas uma possível solução para as deficiências do modelo *bag of words*, mas também uma análise qualitativa das coleções de texto. Os tópicos extraídos pelos modelos servem como resumo dos assuntos presentes na coleção. No lugar da alta dimensionalidade do modelo *bag of words*, os modelos de tópicos representam documentos em baixa dimensionalidade de forma inteligível ao usuário.

1.4 Objetivo e organização

Este documento descreve estudo conduzido sobre a aplicação de *modelos de tópicos* na representação e classificação de resenhas de usuário em relação aos sentimentos expressos e direcionados a múltiplos aspectos do objeto avaliado.

O aumento da acurácia da tarefa de classificação de sentimentos é buscado através da representação de documentos de texto estimada pelo modelo *Latent Dirichlet Allocation* e por quatro variantes. A primeira variante, o modelo de digramas, permite a inclusão de informação posicional na estimação da representação de tópicos. A segunda variante, o modelo de temas, permite a inclusão de informação categórica de documentos. Por fim, dois novos modelos desenvolvidos neste trabalho, o modelo Nota-Aspecto e modelo Nota-Aspecto LDA, são analisados no que se refere à inclusão de avaliações (*ratings*) e à identificação dos tópicos estimados.

São relatados pequenos ganhos na tarefa de classificação de sentimentos com resenhas extraídas de um sítio da internet. Análises qualitativas das representações obtidas com os diferentes modelos de tópicos empregados também são apresentadas.

No Capítulo 2 são apresentadas e discutidas as principais abordagens ao problema de classificação de sentimentos. O Capítulo 3 a abordagem dos modelos de tópicos, e apresenta os algoritmos de estimação de parâmetros implementados. Os dois modelos desenvolvidos pelo autor são descritas no Capítulo 4. Os resultados experimentais obtidos são descritos e discutidos no Capítulo 5. Por fim, as conclusões, discussões sobre o estudo e possíveis melhorias são apresentados no Capítulo 6.

2 CLASSIFICAÇÃO DE SENTIMENTOS

De maneira geral, pode-se dividir as abordagens à classificação de sentimentos encontradas na literatura em duas classes: as baseadas em *recursos linguísticos* e as baseadas em técnicas de *aprendizado de máquina*. A Seção 2.1 apresenta as métricas mais comumente utilizadas na avaliação de tarefas de classificação de sentimentos. As Seções 2.2 e 2.3 detalham as abordagens, com ênfase na abordagem baseada em aprendizado de máquina, foco deste trabalho, ressaltando resultados obtidos em trabalhos anteriores.

2.1 Métricas de avaliação

As técnicas de classificação de sentimentos são geralmente avaliadas através de métricas empregadas nas tarefas de categorização de textos, tais como *acurácia*, *precisão*, *revocação* e *medida F*.

Seja K o número de categorias, a *matriz de confusão* M de um classificador em um conjunto de dados é uma matriz $K \times K$ cujos elementos $M_{i,j}$ são dados pelo número de vezes que um documento pertencente à categoria i é classificado como j. Na matriz, as diagonais $M_{i,i}$ indicam os acertos, e os demais elementos indicam os erros cometidos pelo classificador.

As medidas de *precisão* e *revocação* foram primeiramente utilizadas na tarefa de recuperação de informação para medir, respectivamente, a qualidade e a abrangência do conjunto de documentos retornados para uma busca (MANNING; RAGHAVAN; SCHÜTZE, 2008). Na tarefa de classificação elas indicam, respectivamente, a qualidade e a abrangência em relação a uma categoria i, e são dadas pelas equações

$$\text{precisão}(i) = \frac{M_{i,i}}{\sum_{j=1}^{K} M_{j,i}} \tag{2.1}$$

$$\text{revocação}(i) = \frac{M_{i,i}}{\sum_{j=1}^{K} M_{i,j}}\,. \tag{2.2}$$

Essas duas medidas podem ser resumidas pela sua média harmônica, conhecida como medida F e dada por

$$F(i) = \frac{\text{precisão}(i) \times \text{revocação}(i)}{\text{precisão}(i) + \text{revocação}(i)}. \tag{2.3}$$

A acurácia de um classificador avalia o sistema de maneira geral, ou seja, considerando todas as categorias. Ela é dada por

$$\text{acurácia} = \frac{\sum_{i=1}^{K} M_{i,i}}{\sum_{i=1}^{K} \sum_{j=1}^{K} M_{i,j}}\,. \tag{2.4}$$

2.2 Classificação de sentimentos baseada em recursos linguísticos

A abordagem baseada em recursos linguísticos utiliza bases de conhecimentos linguísticos compiladas para determinar o sentimento de um texto (frase, sentença ou documento), como por exemplo dicionários, redes semânticas, ontologias de linguagem e resultados de mecanismos de busca (DING; LIU; YU, 2008).

Talvez o trabalho mais significativo dentro dessa abordagem seja o de Turney (2001), no qual utiliza-se a informação mútua (IM)[1] entre um texto e discriminadores de sentimentos (palavras como bom, excelente, mal e ruim) para determinar a *polaridade* de uma sentença, isto é, para decidir se a sentença exprime uma opinião positiva, negativa ou neutra. O autor propõe uma métrica baseada na frequência relativa de páginas encontradas para buscas contendo expressões conjuntivas do texto e um dos discriminadores como aproximação para a informação mútua. A polaridade P de uma sentença s é então dada pela equação $P(s) = \text{IM}(s|\text{pos}) - \text{IM}(s|\text{neg})$, onde pos e neg indicam discriminadores relativos a opiniões positivas e negativas, respectivamente. Por fim, a polaridade de um documento é definida pela média das polaridades de suas sentenças. Acurácias entre 66%–84% são relatadas para conjuntos de resenhas de usuários que avaliam diferentes tipos de produtos.

Uma desvantagem das técnicas baseadas em recursos linguísticos é que tais recursos são muito comumente de difíceis obtenção e atualização. No caso de dicionários ou ontologias, eles devem ser fabricados levando-se em conta o domínio da aplicação. No caso de buscadores de internet, eles requerem acesso *online* ao sistema ou uma cópia local, o que normalmente é impraticável devido aos recursos computacionais necessários.

[1] A informação mútua entre dois eventos A e B é dada por $\text{IM} = \log p(A, B) - \log p(A) - \log p(B)$.

2.3 Aprendizado de máquina na classificação de sentimentos

Assim como na tarefa de classificação de textos em tópicos, a maior parte dos trabalhos em classificação de sentimentos utiliza-se de técnicas de aprendizado de máquina para construir classificadores (PANG; LEE, 2008). Entende-se por *classificador* uma função $f : \mathcal{A} \rightarrow \mathcal{C}$, que mapeia documentos representados em um espaço de atributos $\mathcal{A}$ em um espaço de categorias $\mathcal{C}$ (SEBASTIANI, 2002). Na tarefa em questão, o espaço de categorias $\mathcal{C}$ representa os possíveis sentimentos expressos em um documento a serem identificados pelo algoritmo de classificação de sentimentos. Por *aprendizado de máquina*, entendem-se as técnicas de estimação da função f a partir de um conjunto de dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2001).

Os três classificadores relatados na literatura como mais bem sucedidos na tarefa de categorização de texto são: Naive Bayes, Máxima Entropia e SVM (MANNING; SCHÜTZE, 1999; PANG; LEE; VAITHYANATHAN, 2002; SEBASTIANI, 2002). Eles assumem como entrada um documento representado vetorialmente por $d = \langle \phi(a_1), \ldots, \phi(a_m) \rangle$, onde a_i denotam atributos determinados previamente, como por exemplo palavras em um vocabulário, e $\phi : \mathcal{A} \rightarrow \mathbb{R}$ funções reais de atributos, como por exemplo o número de ocorrências do atributo a_i ou uma função binária indicando a presença ou ausência do atributo a_i.

O algoritmo *Naive Bayes* é um dos mais simples classificadores e, no entanto, tem-se mostrado muito eficiente na classificação de dados com alta dimensionalidade, como é o caso de documentos de texto (DOMINGOS; PAZZANI, 1997). A função de decisão escolhe a moda da distribuição de categorias *a posteriori* condicionada nos atributos do documento

$$f_{NB}(d) = \arg\max_c P(c) \prod_{i=1}^{m} P(a_i|c)^{\phi(a_i)}, \tag{2.5}$$

onde a função ϕ comumente denota o número de ocorrências do atributo a_i no documento.

A função f_{NB} pode ser facilmente estimada a partir de um conjunto de dados utilizando-se estimadores de máxima verossimilhança que fornecem uma solução de forma fechada (MANNING; SCHÜTZE, 1999).

O modelo probabilístico do *Naive Bayes* assume que os atributos são condicionalmente independentes. Por exemplo, quando usado na categorização de texto utilizando palavras como atributos, essa independência assume a idéia de que diante da informação categórica de um documento, a ocorrência de uma palavra não influencia na ocorrência de outra.

Essa hipótese é dificilmente verificada em dados reais.

O modelo de Máxima Entropia (MaxEnt) é um modelo probabilístico que tenta superar a hipótese simplista de independência feita pelo modelo *Naive Bayes*. Sua função de decisão é dada por

$$f_{ME}(d) = \arg\max_c \frac{1}{Z(d)} \exp\left(\sum_i \omega_{k,c}\lambda_k(d,c)\right), \tag{2.6}$$

onde $Z(d)$ é uma função de normalização e λ_k são funções positivas que relacionam funções de atributos $\phi(a_i)$ e classes c, e usualmente dadas por

$$\lambda_k(d,c') = \begin{cases} \phi(a_i), & \text{se } \phi(a_i) > 0 \text{ e } c' = c \\ 0, & \text{caso contrário.} \end{cases}$$

Os pesos $\omega_{k,c}$ são parâmetros que indicam a influência de cada função atributo na discriminação de uma categoria c. Eles são estimados encontrando-se valores que maximizem a entropia da distribuição *a posteriori* dada pela função f_{ME} sob as restrição de que os valores esperados das funções atributos sejam iguais aos valores esperados empíricos calculados a partir dos dados (NIGAM, 1999; MALOUF, 2002).

Diferentemente dos outros dois classificadores citados, o classificador *Support Vector Machine* (SVM) não assume um modelo probabilístico. Sua estimação consiste em encontrar o hiperplano representado por um vetor $\mathbf{w}$ que separa os documentos de categorias distintas deixando a maior "margem" ou distância. O vetor $\mathbf{w}$ pode ser encontrado utilizando-se algoritmos de otimização (JOACHIMS, 1999). A função de decisão consiste simplesmente em classificar um documento de acordo com a parte do hiperplano a que ele pertence. O classificador SVM é relatado na literatura de categorização de texto como um dos mais eficazes (JOACHIMS, 1998).

Pang, Lee e Vaithyanathan (2002) apresentam uma extensa análise do problema e da aplicação dos métodos de aprendizagem de máquina à classificação de sentimentos. Os autores propõem identificar o grau de dificuldade da tarefa analisando a capacidade humana de discriminar o sentimento geral expresso no texto. Dois avaliadores são recrutados para classificar um conjunto de dados homogeneamente distribuído (i.e., 50% das resenhas são positivamente avaliadas), e acurácias entre 58% e 64% são relatadas. Diversas configurações de representações de documentos são então empiricamente avaliadas através da determinação de polaridade utilizando-se os classificadores *Naive Bayes*, SVM e de Máxima Entropia. Eles relatam desempenho máximo de $82,9\%$ para um conjunto de atributos contendo unigramas apenas (i.e., um modelo *bag of words*), utilizando-se o classificador

SVM. Representações mais complexas como digramas (pares de palavras), unigramas mais etiquetas sintáticas, unigramas mais posicionamento e uma lista de palavras-chave compilada por especialistas mostraram-se menos eficientes. Os autores concluem observando que as acurácias relatadas evidenciam a maior dificuldade da tarefa de classificação de sentimento em relação à classificação em tópicos, onde acurácias superiores a 90% são facilmente obtidas.

Dave, Lawrence e Pennock (2003) experimentam a incorporação etiquetas sintáticas e reduções morfológicas para atingir desempenhos de até $88,9\%$ para um conjunto de dados de resenhas altamente tendencioso (com 5 vezes mais resenhas positivas que negativas), e até $85,8\%$ em um conjunto de dados equilibrado (com 50% das resenhas rotuladas positivamente). Os resultados obtidos demonstram ligeira superioridade do classificador SVM em relação ao *Naive Bayes*.

O trabalho de Yu e Hatzivassiloglou (2003) divide a tarefa em duas etapas. Primeiro, eles utilizam um classificador *Naive Bayes* para classificar artigos de jornal em opinativos ou fatuais. Eles reportam uma medida F de 96%, que relatam ser compatível com acurácias de 93% obtidas em trabalhos anteriores. Os autores então utilizam o conjunto automaticamente rotulado de artigos para separar sentenças opinativas de sentenças fatuais e determinam a polaridade das sentenças opinativas utilizando um procedimento semelhante ao de Turney (2001). Acurácias entre 37% e 90% são obtidas para diferentes conjuntos de atributos, com o melhor desempenho ocorrendo para um conjunto de atributos de unigramas onde os substantivos foram removidos.

Pang (2005) estende a descrição de sentimentos, permitindo intervalos ao invés de sentimentos binários. Utilizando o classificador *SVM*, documentos são discriminados em quatro classes (0–3), denotando o grau do sentimento expresso no texto. Acurácias de no máximo 66% são obtidas, indicando a maior dificuldade desta tarefa. Os autores então realizam experimentos com avaliadores humanos, indicando que uma descrição mais detalhada dos sentimentos é, de fato, uma tarefa mais árdua, mesmo para humanos.

Snyder e Barzilay (2007) introduzem o conceito de classificação multiaspecto (discutido no Capítulo 1). Eles utilizam um algoritmo do tipo *perceptron* para determinar o sentimento de resenhas numa escala de 1–5. Observando empiricamente que existe uma forte correlação entre as notas dadas a diferentes aspectos em uma mesma resenha, eles desenvolvem um classificador de concordância que decide, baseado nos dados do texto, quando todas as notas terão o mesmo valor, atingindo acurácias de até 67% na tarefa de determinação de concordância. Eles então constroem um meta classificador capaz de

incorporar as informações do classificador de sentimento de cada aspecto com o classificador de concordância, e demonstram que o modelamento explícito do correlacionamento entre aspectos ajuda a melhorar o desempenho dos preditores. Eles relatam erros de ranqueamento médio de $63,20\%$ no melhor caso.[2]

Blei e McAuliffe (2008) estendem o tradicional modelo de tópicos *Latent Dirichlet Allocation* para incluir variáveis de resposta, a serem inferidas a partir de uma regressão sobre os tópicos inferidos. Eles realizam testes com o mesmo conjunto de dados utilizado em (PANG, 2005) e relatam melhoras de quase 10% na métrica preditiva de R^2, definida como 1 - $\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$, onde y_i e $\hat{y}$ denotam os valores verdadeiro e estimado da i-ésima instância no atual conjunto de teste, e $\bar{y}$ denota o valor estimado médio calculado sobre todos os conjuntos de teste.

Mais recentemente, Titov e McDonald (2008b) desenvolveram um modelo de tópicos no qual palavras são amostradas a partir de distribuições locais relativas aos muitos aspectos opinados no texto ou de distribuições globais referentes a assuntos mais gerais. Documentos são representados como compostos de atributos binários definidos pela ocorrência de cada palavra e do rótulo mais proeminente na frase. Utilizando o mesmo classificador que Snyder e Barzilay (2007), eles reportam ganhos de cerca de 5% na função de perda de ranqueamento utilizando os dados do modelo aprendido em relação ao modelo original. Em (TITOV; MCDONALD, 2008a), eles estendem o modelo de forma semelhante ao trabalho de Blei e McAuliffe (2008), permitindo a incorporação de notas de avaliação multiaspecto na forma de regressores. No entanto, eles utilizam o modelo apenas para identificar trechos onde opiniões sobre um aspecto alvo ocorrem.

Mauá e Cozman (2009) utilizam Lógica de Markov (RICHARDSON; DOMINGOS, 2006) para aprender um classificador conjunto de Máxima Entropia capaz de classificar sentenças em múltiplos aspectos. Eles obtêm um conjunto de sentenças rotuladas aplicando técnicas de recuperação de informação. Para cada aspecto, uma busca formada por palavras previamente selecionadas por um dos autores é realizada e as 500 sentenças mais relevantes são rotuladas com o aspecto e adicionadas ao conjunto de treinamento. Para aumentar a revocação das sentenças retornadas, é aplicada uma técnica de redução de dimensionalidade conhecida com *Latent Semantic Indexing* (MANNING; SCHÜTZE, 1999; MANNING; RAGHAVAN; SCHÜTZE, 2008). O classificador de sentenças é então utilizado para rotular as sentenças de um conjunto de teste. Os documentos são representados por vetores cujos elementos são formados pelas palavras rotuladas com o(s) rótulo(s) de

[2] O erro de ranqueamento é dado por $\sum_i \frac{|y_i - \hat{y}_i|}{N}$, onde y_i e $\hat{y}_i$ denotam, respectivamente, as notas real e predita para a i-ésima instância, e N o total de instâncias.

aspecto(s) da sentença a que ela pertença. Uma acurácia média $55,9\%$ é reportada na tarefa de classificação de sentimentos multiaspecto numa escala de 1–5, demonstrando um aumento de cerca de $1,3\%$ sobre a representação *bag of words* usual. Os Exemplos 6 e 7 ilustram o resultado da representação final utilizada no trabalho para as resenhas dos Exemplos 1 e 2, respectivamente. O artigo completo é reproduzido no Anexo B.

Existem três diferenças fundamentais entre o trabalho de Mauá e Cozman (2009) e este. Enquanto que Mauá e Cozman empregam aprendizado supervisionado para anotar as palavras, os modelos de tópicos estudados neste trabalho utilizam um procedimento completamente não supervisionado. Em segundo lugar, os modelos de tópicos atuam no nível das palavras, permitindo que cada palavra seja assinalada a apenas uma categoria. A técnica proposta por Mauá e Cozman, por outro lado, atuam no nível das sentenças, permitindo que cada palavra seja associada a mais de uma categoria. Por fim, enquanto Mauá e Cozman visam apenas o aumento do desempenho da tarefa de classificação de sentimentos, os modelos de tópicos permitem a extração de informação semântica oculta na grande quantidade de dados.

3 MODELOS DE TÓPICOS

Este capítulo dedica-se à discussão do uso de modelos de tópicos para o modelamento de coleção de documentos textuais. O assunto é introduzido na Seção 3.1, com um breve histórico da área. O algoritmo *Latent Dirichlet Allocation* (LDA), o mais comumente utilizado, é então apresentado na Seção 3.2, assim como técnicas para sua estimação. A incorporação de informação posicional ao modelo LDA é apresentada no modelo de digramas LDA na Seção 3.3. Na Seção 3.4 é apresentado o modelo de temas, uma extensão ao modelo LDA original que permite a inclusão de informação categórica comumente disponível na tarefa de classificação de sentimentos. Por fim, na Seção 3.5 são discutidos demais modelos presentes na literatura.

3.1 Análise semântica

Comumente, coleções de textos têm sido representadas nas mais diversas aplicações como matrizes de co-ocorrência de palavras e documentos. Implícita neste tratamento está a hipótese de que a informação contida em um documento pode ser representada pela soma das informações contidas nas palavras que o compõe (MANNING; RAGHAVAN; SCHÜTZE, 2008). Empiricamente, esta abordagem sustenta-se pelos resultados que comprovam que o ganho advindo da incorporação de mais informações na representação dos documentos não supera o prejuízo causado pelo aumento da dimensionalidade dos dados.

Os modelos de análise semântica (HOFMANN, 1999; MANNING; SCHÜTZE, 1999; STEYVERS; GRIFFITHS, 2007; MANNING; RAGHAVAN; SCHÜTZE, 2008) foram desenvolvidos como solução para o impasse entre a incorporação de informação e o aumento de dimensionalidade. O algoritmo de *Análise de Semântica Latente* (do inglês, *Latent Semantic Analysis*, LSA), inclui a noção de que informação semântica pode ser inferida a partir da redução da dimensionalidade do conjunto de dados através de uma decomposição apropriada, e que esta informação de baixa dimensionalidade pode, de fato, representar tão bem ou melhor a coleção, diminuindo consideravelmente os esforços computacionais, e

revelando conceitos "escondidos" nos dados.

Experimentos evidenciaram que a redução da dimensionalidade pode ajudar a aumentar o desempenho de tarefas como categorização de texto e recuperação de informações (MANNING; SCHÜTZE, 1999; MANNING; RAGHAVAN; SCHÜTZE, 2008). Por trás desse ganho informacional estão os fenômenos de polissemia, que faz com que a informação contida em uma palavra leve a ambigüidades, e de sinonímia, que leva a um aumento da dimensionalidade sem efetivo aumento informacional, tornando os dados mais esparsos. Infelizmente, devido ao grande custo computacional da decomposição de matrizes realizada no algoritmo LSA, os bons resultados não refletiram em sua ampla adoção. Ademais, o algoritmo LSA tem por trás de sua formulação hipóteses estatísticas reconhecidamente equivocadas, como admitir que a freqüência de ocorrência de palavras é modelada por uma distribuição normal.[1]

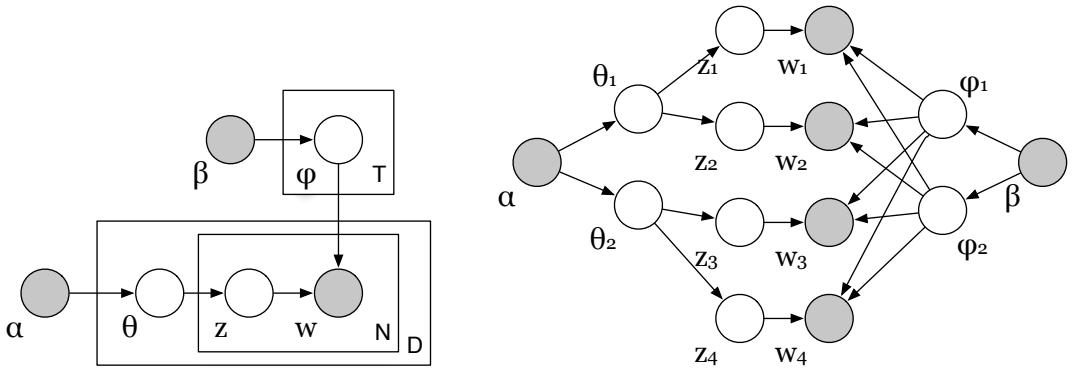
Apoiando-se na idéia de que é possível encontrar uma representação de baixa dimensionalidade que represente mais fielmente as informações contidas em um documento, Hofmann (1999) apresentou uma versão probabilística de um algoritmo de análise semântica chamada de *Probabilistic Latent Semantic Indexing* (pLSI). Diferentemente do algoritmo LSA, o algoritmo pLSI representa um documento como uma mistura de distribuições sobre vocábulos (palavras distintas), onde cada distribuição, denominada *tópico*, representa um possível tema ou assunto.[2]

Um modelo de tópicos é um *modelo generativo*, ou seja, é um modelo especificado por um procedimento probabilístico pelo qual os dados são gerados. O modelo então deve ser estimado encontrando-se o melhor conjunto de variáveis latentes (i.e., não observáveis) que descreve os dados (STEYVERS; GRIFFITHS, 2007). Os modelos de tópicos possuem muitas vantagens sobre os modelos matriciais como o LSA. Diferentemente dos modelos matriciais, os tópicos estimados são facilmente interpretáveis. Ademais, os modelos probabilísticos apresentam maior facilidade de compreensão e de extensão.

O modelo pLSI possui problemas de generalização que podem levar ao "ovefitting" dos parâmetros estimados. Ademais, o pLSI em sua derivação original não permite a estimação de novos documentos. De forma a evitar esses contratempos, Blei et al. (2003) estenderam o modelo pLSI, acrescentando distribuições Dirichlet *a priori* às variáveis de mistura de tópicos e de tópicos, chegando ao modelo *Latent Dirichlet Allocation*.

[1] Como notado por Manning e Schütze (1999), misturas de distribuições de Poisson são reportadas como melhor modelar a ocorrência de palavras.

[2] No artigo original, o autor usa o termo *aspecto* para definir as distribuições sobre vocábulos referentes a um assunto. No entanto, o termo *tópico* foi consagrado pela literatura posterior.



(a) (b)

Figura 2: Representação gráfica do modelo LDA. (a) Representação condensada. Os retângulos indicam repetições das variáveis aleatórias pelo número indicado no canto direito inferior. (b) Representação expandida para uma coleção de dois documentos, quatro palavras e dois tópicos.

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation é um poderoso arcabouço para o modelamento de coleções de dados de contagem que recentemente tem sido aplicado a diversas tarefas, especialmente nas áreas de processamento de linguagem natural e visão computacional (LI; PERONA, 2005; XING, 2007; BHATTACHARYA; GETOOR, 2006; SIVIC et al., 2005).

De forma sucinta, pode-se descrever o processo generativo de um documento descrito pelo LDA da seguinte forma. Primeiro, amostre T vetores positivos φ_j de uma distribuição Dirichlet com parâmetros β. Os vetores φ_j são parâmetros de distribuições discretas de palavras denominadas *tópicos*, que representam os diferentes possíveis assuntos que um autor pode utilizar ao redigir um documento. Então, amostre um vetor positivo θ de uma distribuição Dirichlet com parâmetro α. O vetor θ_d é o parâmetro de uma distribuição discreta que indica a proporção de cada tópico φ_j no documento. Para cada palavra no documento, primeiro amostre uma variável de rótulo de tópico z de uma distribuição discreta com parâmetros θ_d, então amostre uma palavra w do tópico respectivo φ_z. O modelo é representado graficamente na Figura 2(a). No grafo direcionado da figura, cada vértice indica uma variável, e arcos indicam dependências diretas. As caixas indicam repetição de uma variável pelo número no canto inferior direito da caixa. Vértices sombreados e não sombreados representam, respectivamente, variáveis observadas (ou definidas pelo usuário) e latentes. Por exemplo, a Figura 2(b) ilustra o modelo LDA para uma coleção de dois documentos, cada qual com duas palavras, e dois tópicos.

O processo generativo pode ser concisamente descrito pela notação abaixo, onde o

símbolo $\sim$ significa amostrado de.

$$\begin{aligned} \varphi_j &\sim \text{Dirichlet}(\beta) \\ \theta_d &\sim \text{Dirichlet}(\alpha) \\ z_i &\sim \text{Discreta}(\theta_{d_i}) \\ w_i|z_i = j &\sim \text{Discreta}(\varphi_j) \end{aligned}$$

3.2.1 Estimação de parâmetros

As variáveis de rótulo z de cada palavra podem ser estimadas através da probabilidade posterior conjunta[3]

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w},\mathbf{z})}{P(\mathbf{w})} \, . \tag{3.1}$$

A distribuição conjunta de palavras e rótulos $P(\mathbf{w},\mathbf{z})$ pode ser fatorizada, marginalizando-se nas variáveis φ e θ

$$\begin{aligned} P(\mathbf{w},\mathbf{z}) &= P(\mathbf{w}|\mathbf{z})P(\mathbf{z}) \\ &= \left(\int_\varphi P(\mathbf{w}|\mathbf{z},\varphi)P(\varphi)\,d\varphi\right)\left(\int_\theta P(\mathbf{z}|\theta)P(\theta)\,d\theta\right) \, . \end{aligned} \tag{3.2}$$

O primeiro termo do lado direito da Equação corresponde a verosimilhança dos dados condicionada apenas nos rótulos $\mathbf{z}$. Efetuando-se a integral analiticamente chega-se a (GRIFFITHS; STEYVERS, 2004)

$$\begin{aligned} P(\mathbf{w}|\mathbf{z}) &= \int_\varphi P(\varphi)P(\mathbf{w}|\mathbf{z},\varphi)\,d\varphi \\ &= \left(\frac{\Gamma(\sum_k \beta_k)}{\prod_{k=1}^W \Gamma(\beta_k)}\right)^T \prod_{j=1}^T \frac{\prod_{k=1}^W \Gamma(n_j^{(k)}+\beta_k)}{\Gamma(\sum_k n_j^{(k)}+\beta_k)} \, , \end{aligned} \tag{3.3}$$

onde $\Gamma(\cdot)$ denota a função matemática Gama (SEBAH; GOURDON, 2002), W o número de palavras distintas na coleção (vocábulos) e $n_j^{(k)}$ o número de palavras iguais ao k-ésimo vocábulo e atribuídas ao j-ésimo tópico na coleção.

Da mesma forma, calculando-se a integral no segundo termo do lado direito da Equação (3.2) tem-se

$$\begin{aligned} P(\mathbf{z}) &= \int_\theta P(\theta|\alpha)P(\mathbf{z}|\theta)\,d\theta \\ &= \left(\frac{\Gamma(\sum_j \alpha_j)}{\prod_{j=1}^T \Gamma(\alpha_j)}\right)^D \prod_{d=1}^D \frac{\prod_{j=1}^T \Gamma(n_j^{(d)}+\alpha_j)}{\Gamma(\sum_j n_j^{(d)}+\alpha_j)} \, , \end{aligned} \tag{3.4}$$

onde $n_j^{(d)}$ denota o número de palavras atribuídas ao j-ésimo tópico no d-ésimo documento

[3]Exceto quando explícito o contrário, assumem-se todas as distribuições apresentadas aqui condicionadas nas variáveis α e β.

e D o total de documentos na coleção.

Juntando-se as Equações (3.3) e (3.4), tem-se

$$P(\mathbf{w},\mathbf{z}) = AB \prod_{j=1}^{T} \frac{\prod_{k=1}^{W} \Gamma(n_j^{(k)} + \beta_k)}{\Gamma(\sum_k n_j^{(k)} + \beta_k)} \prod_{d=1}^{D} \frac{\prod_{j=1}^{T} \Gamma(n_j^{(d)} + \alpha_j)}{\Gamma(N_d + \alpha_0)} , \tag{3.5}$$

onde

$$A = \left(\frac{\Gamma(\sum_j \alpha_j)}{\prod_{j=1}^{T} \Gamma(\alpha_j)} \right)^D \quad \text{e} \quad B = \left(\frac{\Gamma(\sum_k \beta_k)}{\prod_{k=1}^{W} \Gamma(\beta_k)} \right)^T$$

são constantes dependentes apenas dos hiperparâmetros α e β.

O denominador da Equação (3.1) pode ser obtido marginalizando as variáveis de rótulo $\mathbf{z}$ na Equação (3.2).

$$P(\mathbf{w}) = \int P(\mathbf{w},\mathbf{z})\, d\mathbf{z} \tag{3.6}$$

Infelizmente, a soma na Equação (3.6) envolve o cálculo de T^N termos, onde N é o número de palavras na coleção, o que é impraticável mesmo para conjuntos de dados pequenos. De forma a evitar esta complexidade, métodos de inferência aproximados podem ser utilizados como algoritmos variacionais, *Expectation-Propagation* e *Expectation-Maximization* (MINKA; LAFFERTY, 2002; BLEI et al., 2003; ASUNCION et al., 2009).

Seguindo o trabalho de Griffiths e Steyvers (2004), optou-se por implementar um algoritmo de Gibbs Sampling, relatado como capaz de obter resultados competitivos aos outros métodos em termos computacionais, com as vantagens de facilidade de entendimento e implementação. As equações para a inferência são exibidas na Subseção 3.2.2 a seguir.

3.2.2 Gibbs Sampling

Os métodos de Monte Carlo foram desenvolvidos para a aproximar funções em domínios complexos ou de alta dimensionalidade a partir da média de suas amostras.[4] De forma sucinta os métodos de Monte Carlo realizam a aproximação (ANDRIEU et al., 2003)

$$\int_{\mathcal{X}} f(\mathbf{x}) P(\mathbf{x})\, d\mathbf{x} \approx \frac{1}{R} \sum_{r=1}^{R} f(\mathbf{x}_r) ,$$

onde as amostras x_r são retiradas de uma distribuição $P(\mathbf{x})$.

A questão principal com métodos de Monte Carlo é como obter amostras de distribuições arbitrárias. Os métodos de Cadeias de Markov de Monte Carlo solucionam

[4] O termo *amostra* é utilizado aqui para descrever um indivíduo de uma população.

o problema beneficiando-se de propriedades interessantes das Cadeias de Markov para atingir uma distribuição final alvo a partir de uma distribuição inicial qualquer (GILKS, 1995).

O algoritmo de Gibbs Sampling é um algoritmo de Cadeias de Markov de Monte Carlo que consiste em, a cada passo, amostrar da distribuição de cada variável condicionada em todas as outras. Dessa forma, a fim de calcular $P(\mathbf{z}|\mathbf{w})$, o algoritmo requer amostras de

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})P(z_i = j|\mathbf{z}_{-i}) \ , \tag{3.7}$$

onde a notação $\mathbf{x}_{-i}$ denota o conjunto de todas as variáveis x_k com $k \neq i$.

Integrando a Equação (3.7) nas variáveis θ e φ, e manipulando algebricamente, obtemos

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \frac{(n^{(k)}_{-i,j} + \beta_k)(n^{(d)}_{-i,j} + \alpha_j)}{\sum_k n^{(k)}_{-i,j} + \beta_k} \ , \tag{3.8}$$

onde $n^{(k)}_{-i,j}$ denota o número de palavras $w_{i'} = k$ rotuladas como $z_{i'} = j$ em toda a coleção sem considerar a i-ésima palavra e $n^{(d)}_{-i,j}$ denota o número de palavras rotuladas como j no d-ésimo documento sem considerar a i-ésima palavra.

O valor exato da distribuição posterior marginal pode ser calculado normalizando a Equação (3.8). A derivação completa do algoritmo encontra-se no Apêndice A.

A partir das equações posteriores condicionais em mão, o algoritmo pode facilmente ser implementado. O Algoritmo 1 descreve o processo. Primeiro, as variáveis z são inicializadas com valores arbitrários (linha 1) e as estatísticas suficientes são calculadas (linha 2). Então nas linhas 3–16, o algoritmo é executado por um número arbitrário de interações, escolhido de forma a garantir uma quantidade estatisticamente suficiente de amostras. Para cada palavra no documento o algoritmo calcula o seu índice no vocabulário (linha 6) e então amostra um rótulo z a partir da função de distribuição cumulativa de mistura de tópicos do documento (linhas 7–13). Na linha 14, as estatísticas são atualizadas com os novos valores das amostras. Para garantir convergência, as amostras são coletadas após um intervalo inicial denominado *burn in* (linhas 15–16). A cada iteração, o algoritmo possui complexidade $O(T(W+D))$ para armazenamento dos parâmetros. A complexidade computacional por iteração é de $O(NT)$. Uma prática comum não detalhada no algoritmo consiste em estabelecer um intervalo entre a coleta de duas amostras consecutivas, de forma a evitar correlação entre as amostras.

Em qualquer iteração do algoritmo é possível estimar os tópicos φ e as distribuições de tópicos θ_d de cada documento, respectivamente, pelos valores esperados condicionados

Entrada: D documentos com N_d palavras cada
Saída : Amostras de rótulos de tópicos z para cada palavra

1 inicializar cada z_i com um valor aleatório em $[1, T]$
2 calcular estatísticas $n_j^{(k)}$ e $n_j^{(d)}$
3 **para** $iter \leftarrow 1$ **até** NumIterações **faça**
4 **para** $d \leftarrow 1$ **até** D **faça**
5 **para** $i \leftarrow 1$ **até** N_d **faça**
6 $k \leftarrow \{l | v_l = w_i, v_l \in \text{vocabulário}\}$
7 $p_0 \leftarrow 0$
8 **para** $j \leftarrow 1$ **até** T **faça**
9 $p_j \leftarrow p_{j-1} + \frac{(n_{-i,j}^{(k)} + \beta_k)(n_{-i,j}^{(d)} + \alpha_j)}{\sum_k n_{-i,j}^{(k)} + \beta_k}$
10 $u \leftarrow$ amostra de Uniforme$[0, 1]$
11 $z_i \leftarrow 1$
12 **enquanto** $u > p_j / p_T$ **faça**
13 $z_i \leftarrow z_i + 1$
14 atualizar estatísticas $n_j^{(k)}$ e $n_j^{(d)}$
15 **se** $iter >$ BurnIn **então**
16 armazenar z

Algoritmo 1: Algoritmo Gibbs Sampling para o modelo Latent Dirichlet Model.

nas amostras $\mathbf{z}$

$$\hat{\varphi}_j^{(k)} = E[\varphi_j^{(k)} | \mathbf{w}, \mathbf{z}] = \frac{n_j^{(k)} + \beta_k}{\sum_k n_j^{(k)} + \beta_k} \tag{3.9}$$

e

$$\hat{\theta}_d^{(j)} = E[\theta_d^{(j)} | \mathbf{w}, \mathbf{z}] = \frac{n_j^{(d)} + \alpha_j}{\sum_j n_j^{(d)} + \alpha_j} \, . \tag{3.10}$$

Porteous et al. (2008) descrevem uma técnica para acelerar a amostragem da distribuição posterior realizada nas linhas 7–14 do Algoritmo 1 calculando limites ao fator de normalização, ao invés de computá-lo como a soma de todas as probabilidades não normalizadas. Ganhos significativos de velocidade são relatados para modelos com número de tópicos T grande (>500).

3.2.3 Estimação de hiperparâmetros

Os hiperparâmetros α e β são responsáveis pela suavização das distribuições discretas, ajudando na generalização do modelo aprendido e na diminuição de problemas de "overfitting". Como demonstrado empiricamente por Asuncion et al. (2009), eles podem influir consideravelmente no desempenho.

Os hiperparâmetros podem ser estimados a partir dos dados por uma versão Gibbs Sampling do algoritmo *Expectation-Maximization* (WALLACH, 2006). O Algoritmo 2 apresenta o procedimento em pseudo-código. No passo E, a distribuição $P(\mathbf{w}, \mathbf{z}|\alpha, \beta)$ é aproximada tomando-se a média sobre S estimativas pontuais $\hat{P}(\mathbf{w}, \mathbf{z}^{(s)}|\alpha, \beta)$, calculadas a partir de amostras $\mathbf{z}^{(s)}$ retiradas de $P(\mathbf{z}|\mathbf{w}, \alpha^{(i-1)}, \beta^{(i-1)})$, como descritas na Subseção 3.2.2. Então, cada parâmetro é atualizado tomando-se o valor que maximiza a distribuição estimada $\hat{P}(\mathbf{w}, \mathbf{z}|\alpha, \beta)$.

Entrada: D documentos com N_d palavras cada
Saída : Estimativas dos hiperparâmetros *alpha* e *beta*

1 inicialize $\mathbf{z}^{(0)}$, $\alpha^{(0)}$ e $\beta^{(0)}$ aleatoriamente
2 **para** $i \leftarrow 1$ **to** NumIterações **faça**
  `/* passo E */`
3 Amostre S amostras $\{\mathbf{z}^{(s)}\}_{s=1}^{S}$ de $P(\mathbf{z}|\mathbf{w}, \alpha^{(i-1)}, \beta^{(i-1)})$ (Use o Algoritmo 1)
  `/* passo M */`
4 $\alpha^{(i)} = \arg\max_{\alpha} \frac{1}{S} \sum_{s=1}^{S} \log P(\mathbf{w}, \mathbf{z}^{(s)}|\alpha, \beta)$
5 $\beta^{(i)} = \arg\max_{\beta} \frac{1}{S} \sum_{s=1}^{S} \log P(\mathbf{w}, \mathbf{z}^{(s)}|\alpha, \beta)$

Algoritmo 2: Algoritmo estimação dos hiperparâmetros α e β para o modelo LDA através do procedimento de *Expectation-Maximization* Gibbs Sampling.

Para cada amostra $z^{(s)}$, $\log P(\mathbf{w}, \mathbf{z}^{(s)}|\alpha, \beta)$ pode ser facilmente derivado a partir da Equação (3.5) como

$$\begin{aligned}\log P(\mathbf{w}, \mathbf{z}|\alpha, \beta) = \\ \underbrace{\sum_{d=1}^{D}\left(\sum_{j=1}^{T}\Big(\log\Gamma(n_j^{(d)}+\alpha_j) - \log\Gamma(\alpha_j)\Big) + \log\Gamma(\sum_j \alpha_j) - \log\Gamma(\sum_j n_j^{(d)} + \alpha_j)\right)}_{\log P(\mathbf{z}|\alpha)} \\ + \underbrace{\sum_{j=1}^{T}\left(\sum_{k=1}^{W}\Big(\log\Gamma(n_j^{(k)}+\beta_k) - \log\Gamma(\beta_k)\Big) + \log\Gamma(\sum_k \beta_k) - \log\Gamma(\sum_k n_j^{(k)} + \beta_k)\right)}_{\log P(\mathbf{w}|\mathbf{z}, \beta)}.\end{aligned} \tag{3.11}$$

O passo M é realizado encontrando-se valores para α e β que maximizem a Equação (3.11) acima calculada nos valores da iteração anterior. Infelizmente, não há solução de forma fechada para este problema. No entanto, algoritmos iterativos podem ser aplicados tais como subida de gradiente ou Newton-Rapshod. Utiliza-se aqui um esquema de *ponto fixo* descrito por Minka (2003) e utilizado no modelo LDA de Wallach (2006). A derivação das equações de ponto fixo para a Equação (3.11) encontram-se no Apêndice B.

As equações finais de atualização para a i-ésima iteração são dadas por

$$\alpha_j^{(i)} = \alpha_j^{(i-1)} \frac{\sum_{s=1}^{S} \sum_{d=1}^{D} \Psi(n_j^{(d)} + \alpha_j^{(i-1)}) - \Psi(\alpha_j^{(i-1)})}{\sum_{s=1}^{S} \sum_{d=1}^{D} \Psi(\sum_j n_j^{(d)} + \alpha_j^{(i-1)}) - \Psi(\sum_j \alpha_j^{(i-1)})} \tag{3.12}$$

e

$$\beta_k^{(i)} = \beta_k^{(i-1)} \frac{\sum_{s=1}^{S} \sum_{j=1}^{T} \Psi(n_j^{(k)} + \beta_k^{(i-1)}) - \Psi(\beta_k^{(i-1)})}{\sum_{s=1}^{S} \sum_{j=1}^{t} \Psi(\sum_k n_j^{(k)} + \beta_k^{(i-1)}) - \Psi(\sum_k \beta_k^{(i-1)})} \,, \tag{3.13}$$

onde Ψ é a função digama dada por $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$ (SEBAH; GOURDON, 2002).

Devido ao processo de estimação dos parâmetros por amostragem (passo E), o algoritmo de estimação de hiperparâmetros possui alta complexidade computacional.

3.2.4 Predição

Com os tópicos estimados a partir de um conjunto de dados de *treinamento*, é possível utilizar o modelo para predizer os rótulos de tópicos e as distribuições de misturas de tópicos para novos documentos. Seja $\mathbf{w}_{\text{novo}}$ um novo documento (i.e., um documento não presente na coleção utilizada na estimação de tópicos) com N_{novo} palavras. Denote também por $\mathbf{w}_{\text{antigo}}$ a coleção utilizada na estimação de tópicos, e $\mathbf{z}_{\text{antigo}}$, θ_{antigo} e φ as variáveis estimadas. A distribuição *a posteriori* dos rótulos do novo documento é então

$$P(\mathbf{z}_{\text{novo}}|\mathbf{w}_{\text{novo}}, \mathbf{w}_{\text{antigo}}, \mathbf{z}_{\text{antigo}}, \varphi) = \frac{P(\mathbf{z}_{\text{novo}}, \mathbf{w}_{\text{novo}}|\mathbf{w}_{\text{antigo}}, \mathbf{z}_{\text{antigo}}, \varphi)}{P(\mathbf{w}_{\text{novo}}|\mathbf{w}_{\text{antigo}}, \mathbf{z}_{\text{antigo}}, \varphi)} \,. \tag{3.14}$$

Assim como na Equação (3.5), o denominador no lado direito da Equação (3.14) acima envolve uma soma intratável sobre $T^{N_{\text{novo}}}$ termos. Novamente, podem-se utilizar algoritmos aproximados para a inferência. A equação do algoritmo Gibbs Sampling para amostragem de rótulos condicionados nos tópicos φ estimados é dada por

$$\begin{aligned} P(z_i = j|\mathbf{w}) &\propto P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) P(z_i = j|, \mathbf{z}_{-i}) \\ &\propto \varphi_{j_k}(n_{-i,j} + \alpha_j) \,. \end{aligned} \tag{3.15}$$

3.3 Além de unigramas

Uma premissa básica do modelo LDA é a de que documentos podem ser representados por *bag of words*, ou seja, que a ordem das palavras é irrelevante. Esta premissa é razoável quando se está interessado apenas na estrutura latente da coleção, como por exemplo, na tarefa de extração de tópicos. Entretanto, quando o interesse encontra-se na predição

de rótulos de palavras, o descarte da informação posicional das palavras pode levar a desempenhos insatisfatórios, pois apesar de qualquer permutação de atribuição de valores aos rótulos satisfazer o modelo, apenas uma combinação particular satisfaz o problema de rotulação de palavras. Além do mais, quando a distribuição de tópicos de um documento apresenta um grande desbalanceamento concentrando sua massa em um pequeno conjunto de tópicos, algoritmos de inferência por amostragem como Gibbs Sampling tendem a superestimar a informação da distribuição de tópicos em detrimento da informação de distribuição de palavras na estimação dos rótulos.

Wallach (2006) descreve então uma extensão ao modelo LDA que permite a inclusão de digramas (i.e., pares de palavras), numa tentativa de superar as falhas do modelo tradicional transformando tópicos em distribuições de palavras condicionais. Dessa forma, o modelo assume que um tópico φ_j é um vetor de tamanho W^2 que mapeia pares de palavras adjuntas (w_i, w_{i-1}) em probabilidades. Ao invés de contar com T distribuições de palavras como no modelo LDA, o modelo de digramas contém WT distribuições, aumentando os requerimentos de memória. As estatísticas suficientes para o modelo são dadas por $n_j^{(k,k')}$, que denota o número de palavras $w_i = v_k$ precedidas por palavras $w_{i-1} = v_{k'}$ assinaladas ao j-ésimo tópico, e $n_j^{(d)}$, que a exemplo do modelo original denota o número de palavras assinaladas ao j-ésimo tópico no d-ésimo documento.

O processo generativo pode ser concisamente descrito pela notação abaixo. Note que a única alteração em relação ao modelo LDA original é a segunda variável no índice dos tópicos.

$$\begin{aligned}
\varphi_{(j,k)} &\sim \text{Dirichlet}(\beta) \\
\theta_d &\sim \text{Dirichlet}(\alpha) \\
z_i &\sim \text{Discreta}(\theta_{d_i}) \\
w_i | w_{i-1} = k, z_i = j &\sim \text{Discreta}(\varphi_{(j,k)})
\end{aligned}$$

As equações para inferência por Gibbs Sampling são semelhantes às do modelo original, e são dadas por

$$P(z_i = j | \mathbf{z}_{-1}, \mathbf{w}) \propto \frac{n_{-i,j}^{(k,k')} + \beta_k}{\sum_k n_{-i,j}^{(k)} + \beta_k} (n_{-i,j}^{(d)} + \alpha_j) \; . \tag{3.16}$$

As distribuições de tópicos de cada documento θ_d e os tópicos φ_j são estimados de maneira similar ao modelo LDA original,

$$\hat{\varphi}_j^{(k,k')} = E[\varphi_j^{(k,k')} | \mathbf{w}, \mathbf{z}] = \frac{n_j^{(k,k')} + \beta_k}{N_j + \beta_0} \tag{3.17}$$

$$\hat{\theta}_d^{(j)} = E[\theta_d^{(j)}|\mathbf{w}, \mathbf{z}] = \frac{n_j^{(d)} + \alpha_j}{\sum_j n_j^{(d)} + \alpha_j} \,. \tag{3.18}$$

As equações de predição e estimação de hiperparâmetros podem ser obtidas substituindo as estatísticas de tópicos de unigramas $n_j^{(k)}$ por suas contrapartidas de digramas $n_j^{(k,k')}$.

3.4 Incluindo informações categóricas

O modelo LDA provê informações ricas sobre a coleção tais como uma representação de baixa dimensionalidade dos documentos como pontos θ_d no espaço de tópicos e agrupamento das palavras em tópicos φ_j (BLEI; LAFFERTY, 2009). Apesar de sua imensa utilidade, LDA em sua forma original não permite a incorporação de informações extra comumente presentes em documentos.

Recentemente, muitos trabalhos tem sido dedicados ao desenvolvimento de extensões do modelo original de forma a permitir o aproveitamento de informações adicionais como autoria, citações, rótulos de categoria e elementos de metadados genéricos (BLEI; JORDAN, 2003; BLEI; MCAULIFFE, 2008; LI; PERONA, 2005; LACOSTE-JULIEN; SHA; JORDAN, 2008; ROSEN-ZVI et al., 2004).

Resenhas de usuário são normalmente anotadas com avaliações (*ratings*) ou notas que exprimem o sentimento em relação a aspectos do objeto. Essas avaliações podem ser entendidas como informação categórica de documentos e utilizadas para aumentar o desempenho do modelo.

Li e Perona (2005) desenvolvem o modelo de temas, que incorpora informação de categoria de documento através do condicionamento do parâmetro α de suavização da distribuição de tópicos no rótulo c do documento. O procedimento generativo é descrito por

$$\begin{aligned}
\varphi_j &\sim \text{Dirichlet}(\beta) \\
c_d &\sim \text{Discreta}(\eta) \\
\theta_d|c_d = \ell &\sim \text{Dirichlet}(\alpha_\ell) \\
z_i &\sim \text{Discreta}(\theta_d) \\
w_i|z_i = j, \varphi_j &\sim \text{Discreta}(\varphi_j)
\end{aligned}$$

e ilustrado na Figura 3.

Neste modelo, a distribuição a priori Dirichlet global dada pela especificada pela

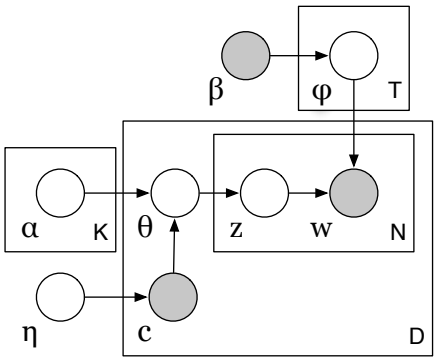


Figura 3: Representação gráfica do modelo de temas.

variável α é transformada em um conjunto de K distribuições especificadas pelas variáveis α_ℓ, uma para cada possível categoria $c \in \{1, \ldots, K\}$. Essa alteração faz com que aumente a probabilidade de documentos de mesma classe possuírem distribuições de tópico θ_D próximas. Note na Figura 3 que a variável α do modelo LDA é substituída por uma coleção de variáveis e que o vértice não é mais sombreado, indicando que essas variáveis devem ser estimadas a partir dos dados. Quando todas as variáveis α_ℓ são iguais entre si, o modelo de temas torna-se o modelo LDA original.

3.4.1 Estimação de parâmetros

As equações para o algoritmo Gibbs Sampling para o modelo de temas são muito semelhantes as do modelo LDA original, diferenciando-se no parâmetro α_ℓ condicionado no rótulo dos documento.

$$P(z_i = j | c_d = \ell, \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{(n^{(k)}_{-i,j} + \beta_k)(n^{(d)}_{-i,j} + \alpha_{\ell_j})}{\sum_k n^{(k)}_{-i,j} + \beta_k} \tag{3.19}$$

3.4.2 Estimação de hiperparâmetros

No modelo de temas, a informação de categoria é incorporada através de K hiperparâmetros α_ℓ no lugar de um único hiperparâmetro utilizado no LDA original. As estimativas de cada elemento α_{ℓ_j} de um parâmetro α_ℓ podem ser obtidas de maneira similar à estimação de hiperparâmetros no modelo LDA, através da seguinte regra de atualização de ponto fixo

$$\alpha^{(i)}_{\ell_j} = \alpha^{(i-1)}_{\ell_j} \frac{\sum_{s=1}^{S} \sum_{d=1}^{D} \mathbf{1}(c_d = \ell) \left(\Psi(n^{(d)}_j + \alpha^{(i-1)}_{\ell_j}) - \Psi(\alpha^{(i-1)}_{\ell_j}) \right)}{\sum_{s=1}^{S} \sum_{d=1}^{D} \mathbf{1}(c_d = \ell) \left(\Psi(\sum_j n^{(d)}_j + \alpha^{(i-1)}_{\ell_j}) - \Psi(\sum_j \alpha^{(i-1)}_{\ell_j}) \right)} \,. \tag{3.20}$$

A funções identidade $\mathbf{1}(\cdot)$ nas somatórias internas indicam que a soma faz-se apenas nos documentos cujos rótulos sejam iguais a ℓ. As equações de atualização para o parâmetro β continuam inalteradas em relação ao modelo original.

O parâmetro η pode ser facilmente estimado por máxima verossimilhança (MV) a partir dos dados de rótulos de classes,

$$\eta_\ell = \arg\max_{\eta_\ell} \log P(\mathbf{c}|\eta) \ . \tag{3.21}$$

Dado η, as variáveis de rótulo tornam-se independentes de seus não descendentes. Dessa forma, a log verossimilhança pode ser decomposta em

$$\log P(\mathbf{c}|\eta) = \sum_{d=1}^{D} \log P(c_d|\eta) = \sum_{\ell=1}^{K} \eta_\ell^{n_\ell} \ , \tag{3.22}$$

onde n_ℓ é o numero de documentos rotulados como ℓ.

As estimativas por MV são obtidas tomando as derivadas parciais do termo mais à direita da Equação (3.22) igual a zero, com a restrição que os elementos do vetor devem somar um, i.e., $\sum_\ell^K \eta_\ell = 1$. A solução em forma fechada é dada por

$$\eta_\ell = \frac{n_\ell}{D}. \tag{3.23}$$

3.4.3 Predição

O algoritmo de Gibbs Sampling pode ser aplicado, marginalizando nas variáveis de rótulos latentes c_d a distribuição marginal de rótulos não condicionada

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto \sum_{\ell=1}^{K} \frac{(n_{-i,j}^{(k)} + \beta_k)(n_{-i,j}^{(d)} + \alpha_{\ell_j})}{\sum_k n_{-i,j}^{(k)} + \beta_k} \eta_\ell \ . \tag{3.24}$$

A cada iteração, pode-se obter estimativas de ML para as variáveis de classe c_d a partir das amostras $\mathbf{z}_d$

$$\begin{aligned}
\hat{c}_d &= \arg\max_{c_d} P(c_d|\mathbf{z}_d, \eta) \\
&= \arg\max_{c_d} P(\mathbf{z}_d|c_d)P(c_d) \\
&= \arg\max_{c_d} P(c_d) \textstyle\int P(\mathbf{z}_d|\theta_d)P(\theta_d|c_d)d\,\theta_d \\
&= \arg\max_{c_d} \log P(c_d) \textstyle\int P(\mathbf{z}_d|\theta_d)P(\theta_d|c_d)d\,\theta_d \\
&= \arg\max_{c_d} \log \eta_d + \log\Gamma(\textstyle\sum_j \alpha_{c_{dj}}) - \log\Gamma(\textstyle\sum_j n_j^{(d)} + \alpha_{c_{dj}}) \\
&\qquad + \textstyle\sum_{j=1}^{T} \left[\log\Gamma(n_j^{(d)} + \alpha_{c_{dj}}) - \log\Gamma(\alpha_{c_{dj}})\right] \ .
\end{aligned} \tag{3.25}$$

3.5 Outros modelos

Inúmeras outras extensões ao modelo LDA original podem ser encontradas na literatura, e existe, atualmente, uma pesquisa muita intensa no desenvolvimento de novas extensões que permitam a incorporação de metadados. Nesta seção são descritas algumas das abordagens existentes.

Blei e Jordan (2003) apresentam uma extensão ao LDA chamada CorrLDA que permite que anotações relacionadas aos rótulos de palavras sejam incorporadas no modelo. Eles aplicam o modelo numa coleção de imagens anotadas com palavras que as descrevem. O modelo então busca encontrar a correlação entre as palavras visuais presentes nas imagens e as palavras encontradas nos textos anotados.

Em (GRIFFITHS et al., 2005), o modelo é incorporado em um Modelo Oculto de Markov, de forma a incorporar informação sintática. Primeiramente, o modelo de Markov é utilizado para separar palavras de função de alta ocorrência no texto e com menor poder semântico, como artigos e conjunções, de palavras com maior "significância". Então um modelo LDA é estimado considerando-se desprezando as palavras de função. Eles reportam melhoras na qualidade dos tópicos estimados.

Como citado anteriormente, Rosen-Zvi et al. (2004) estendem o modelo LDA para incluir informação de autoria de documentos. Eles descrevem um modelo no qual documentos são descritos generativamente como possuindo vários autores. Eles então desenvolvem dois modelos, um no qual os tópicos são dependentes da escolha do autor, e outro no qual existem variáveis de distribuição de tópicos descrevendo a preferência de cada autor por certos tópicos.

Lacoste-Julien, Sha e Jordan (2008) implementam no modelo DiscLDA uma transformada que mapeia as variáveis de distribuições de tópicos de documentos de mesma categoria em pontos próximos, da mesma forma que o modelo de temas. Diferentemente do modelo de temas, as variáveis são mapeadas por uma matriz de transformação cujos parâmetros são aprendidos discriminativamente.

Blei e McAuliffe (2008) incorporam variáveis de resposta no modelo dependentes das variáveis de rótulo de cada documento. Diferentemente das outras abordagens, as variáveis de categoria de documento não afetam diretamente, mas não afetas pelas informações de rótulo e palavras de documentos. As variáveis de classe são então estimadas por regressores.

Mimno e McCallum (2008) propõem um modelo genérico que permite a inclusão de atributos binários relativos a um documento. O modelo DMR proposto assemelha-se ao modelo de temas, pois as informações são utilizadas para alterar os parâmetros de suavização das distribuições de tópicos. No entanto, diferentemente do modelo de tópicos o DMR permite que vários atributos sejam codificados.

4 OS MODELOS NOTA-ASPECTO

Neste capítulo são descritas as duas extensões ao modelo LDA que permitem à inclusão de avaliações (*ratings*) de resenhas e identificam explicitamente os tópicos aprendidos por um par (aspecto, sentimento). As Seções 4.1 e 4.2 apresentam, respectivamente, os modelos Nota-Aspecto e Nota-Aspecto LDA.

4.1 O modelo Nota-Aspecto

O modelo de temas permite a inclusão de apenas uma informação categórica por documento. Entretanto, como é o caso de resenhas avaliadas em relação a muitos aspectos, em muitos domínios existe múltiplas informações categóricas a respeito de um documento. O modelo Nota-Aspecto e sua extensão descrita na Seção 4.2, propostos neste trabalho e baseados nos modelos de tópicos de autoria de Rosen-Zvi et al. (2004), incorporam informações de avaliações de resenhas de uma forma alternativa ao modelo de temas, permitindo a inclusão de múltiplas informações por documento.

O modelo Nota-Aspecto descreve um processo no qual, primeiro, amostram-se $T = A \times K$ tópicos de uma distribuição Dirichlet com parâmetros β, onde A é o número de aspectos e K o número de notas que um usuário pode utilizar para avaliar um objeto em relação a um determinado aspecto. Para cada palavra no documento, escolhe-se um aspecto para comentar amostrando um rótulo a de uma distribuição discreta η comum a todos os documentos. Então, dadas as notas r_a providas pelo usuário no determinado aspecto a, uma palavra w é amostrada do correspondente tópico $\varphi_{(a,r_a)}$. O processo é descrito em notação formal em seguida e representado graficamente na Figura 4.

$$
\begin{aligned}
\varphi_{(j,y)} &\sim \text{Dirichlet}(\beta) \\
r_j^{(d_i)} | a_i = j &\sim \text{Discreta}(\xi_j) \\
a_i &\sim \text{Discreta}(\eta) \\
w_i | a_i = j, r_j^{(d_i)} = y &\sim \text{Discreta}(\varphi_{(j,y)})
\end{aligned}
$$

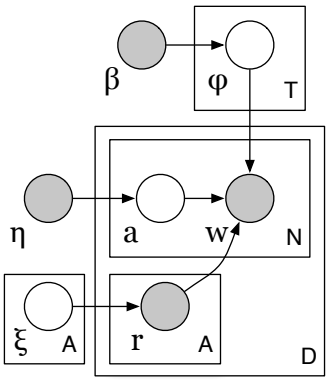


Figura 4: Representação gráfica do modelo Nota-Aspecto.

Uma propriedade interessante deste modelo é que, diferentemente do modelo LDA, os tópicos são diretamente identificáveis, ou seja, é possível determinar rótulos para os tópicos estimados dos dados de maneira automática, pois eles são indexados pelo par aspecto-nota, conhecido *a priori*.

4.1.1 Estimação de parâmetros

Mais uma vez, é possível empregar o algoritmo e Gibbs Sampling para obter amostras da distribuição *a posteriori* de rótulos. As probabilidades marginais são dadas por

$$\begin{aligned} P(a_i = j|\mathbf{a}_{-i}, \mathbf{w}, \mathbf{r}) &\propto P(w_i = k|a_i = j, \mathbf{a}_{-i}, \mathbf{w}_{-i}, \mathbf{r})P(a_i = j) \\ &\propto \frac{n^{(k)}_{-i,(j,y)}+\beta_k}{\sum_k n^{(k)}_{-i,(j,y)}+\beta_k}\eta_j \ , \end{aligned} \tag{4.1}$$

onde o termo $n^{(k)}_{-i,(j,y)}$ denota o número de vezes que a o k-ésimo vocábulo aparece rotulado como aspecto j em um documento onde o aspecto j recebe nota y, sem considerar a i-ésima palavra.

4.1.2 Estimação de hiperparâmetros

A equações de atualização do parâmetro β para o esquema de estimação via ponto fixo são exatamente iguais ao modelo LDA original, levando-se em conta que cada par aspecto, nota (j, y) corresponde a um tópico do modelo LDA.

4.1.3 Predição

De maneira similar ao modelo de temas, as equações para o algoritmo Gibbs Sampling na tarefa de predição podem ser obtidas marginalizando-se a distribuição de rótulos posterior nas variáveis de notas latentes.

$$P(a_i = j|\mathbf{a}_{-i}, \mathbf{w}, \varphi) \propto P(w_i|a_i = j, \mathbf{a}_{-i}, \mathbf{w}_{-i}, \hat{\varphi})P(a_i = j) \ , \tag{4.2}$$

onde a distribuição condicional de palavras é dada por

$$P(w_i|a_i = j, \mathbf{a}_{-i}, \mathbf{w}_{-i}, \varphi) = \sum_{y=1}^{K} P(w_i|a_i = j, r_j^{(d_i)} = y, \varphi_{(j,y)})P(r_j^{(d_i)} = y|\cdot) \ . \tag{4.3}$$

O termo mais à direita na Equação (4.3), relativo à probabilidade condicional da nota referente ao j-ésimo aspecto, pode ser obtido aplicando-se a regra de Bayes

$$\begin{aligned} P(r_j^{(d_i)} = y|a_i = j, \mathbf{a}_{-i}, \mathbf{w}_{-i}^{(d)}, \varphi_{(j,y)}) &= \frac{P(\mathbf{w}_{-i}^{(d_i)}|\mathbf{a}_{-i}^{(d_i)}, r_j^{(d_i)}=y, \varphi)P(r_j^{(d_i)}=y)}{P(\mathbf{w}_{-i}^{(d_i)}|\mathbf{a}_{-i}^{(d_i)}, \varphi)} \\ &= \frac{1}{Z_j}\xi_{j_y}\prod_{k=1}^{W}\varphi_{(j,y)_k}^{n_{-i,j}^{(k,d_i)}} \ , \end{aligned} \tag{4.4}$$

onde $n_{-i,j}^{(k,d)}$ é o numero de palavras k rotuladas como aspecto j no documento d, e Z_j é um fator de normalização dado por $\frac{\sum_{\mathbf{r}_{-j}^{(d_i)}} P(\mathbf{r}_{-j}^{(d_i)}) \prod_{i'} P(w_{i'}|a_{i'} \neq j, \cdot)}{P(\mathbf{w}_{-i}^{(d_i)}|\cdot)}$.

A cada iteração, é possível obter estimativas de MAP para as notas r_j a partir dos rótulos amostrados $\mathbf{a}^{(d)}$ relativos ao d-ésimo documento

$$\begin{aligned} \hat{r}_j^{(d)} &= \arg\max_y P(r_j^{(d)} = y|\mathbf{a}^{(d)}, \mathbf{w}^{(d)}, \varphi) \\ &= \arg\max_y P(\mathbf{w}^{(d)}|\mathbf{a}^{(d)}, r_j^{(d)} = y, \varphi)P(r_j^{(d)}) \\ &= \arg\max_y P(r_j^{(d)}) \prod_{i,a_i=j} P(w_i|a_i = j, r_j^{(d)} = y, \varphi) \\ &= \arg\max_y \log P(r_j^{(d)}) + \sum_{i,a_i=j} \log P(w_i|a_i = j, r_j^{(d)} = y, \varphi) \\ &= \arg\max_y \log \xi_{j_y} + \sum_{k=1}^{W} n_j^{(d)} \log \varphi_{(j,y)}^{(k)} \ . \end{aligned} \tag{4.5}$$

4.2 O modelo Nota-Aspecto LDA

No modelo Nota-Aspecto não há clara indicação da estrutura de documentos senão pelas próprias notas. Entretanto, como é possível notar da Equação (4.1), quando condicionado na informação de nota, qualquer informação específica sobre a estrutura de documentos é completamente perdida, o que pode fazer com que documentos sejam mal representados. No modelo LDA original, tal questão é evitada com a introdução de

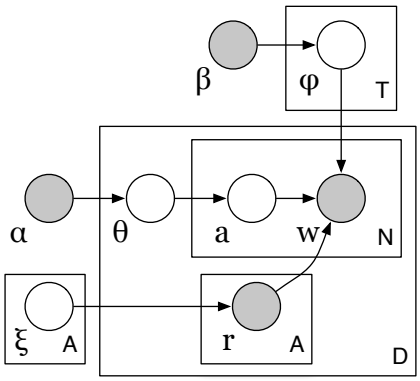


Figura 5: Representação gráfica do modelo Nota-Aspecto LDA.

variáveis Dirichlet que produzem uma representação de baixa dimensionalidade de um documento. O modelo Nota-Aspecto LDA estende o modelo Nota-Aspecto através da incorporação de variáveis Dirichlet sobre os documentos denotando a proporção de aspectos, de maneira similar ao LDA. O processo generativo do modelo é descrito a seguir.

$$\begin{aligned}
\varphi_{(j,y)} &\sim \text{Dirichlet}(\beta) \\
\theta_d &\sim \text{Dirichlet}(\alpha) \\
r_{d_j} &\sim \text{Discreta}(\xi_j) \\
a_i &\sim \text{Discreta}(\theta_d) \\
w_i | a_i = j, r_{d_j=y} &\sim \text{Discreta}(\varphi_{(j,y)})
\end{aligned}$$

O modelo é representado graficamente na Figura 5. A grande diferença em relação ao modelo Nota-Aspecto encontra-se nas distribuições *a priori* das variáveis de aspecto, que no modelo Nota-Aspeto LDA são particulares de cada documento, ao invés de comum a todos os documentos como no modelo Nota-Aspecto.

4.2.1 Estimação de parâmetros

As probabilidades marginais para o algoritmo Gibbs Sampling são dadas por

$$\begin{aligned}
P(a_i = j | \mathbf{a}_{-i}, \mathbf{w}, \mathbf{r}, \beta, \xi) &\propto P(w_i = k | a_i = j, \mathbf{a}_{-i}, \mathbf{w}_{-i}, \mathbf{r}) P(a_i = j | \mathbf{a}_{-i}) \\
&\propto (n^{(d)}_{-i,j} + \alpha_j) \frac{n^{(k)}_{-i,(j,y)} + \beta_k}{\sum_k n^{(k)}_{-i,(j,y)} + \beta_k} .
\end{aligned} \tag{4.6}$$

Como é possível notar na Equação (4.6) acima, as variáveis de probabilidade *a priori* η_l da Equação (4.1) do modelo Nota-Aspecto são substituídas no modelo Nota-Aspecto LDA pela componente $(n^{(d)}_{-i,j} + \alpha_j)$, que inclui uma contagem de atribuições por documento.

4.2.2 Estimação de hiperparâmetros

Assim como no modelo Nota-Aspecto as equações de atualização dos hiperparâmetros para o esquema de estimação via ponto fixo são exatamente iguais ao modelo LDA original, levando-se em conta que cada par aspecto, nota (j, y) corresponde a um tópico do modelo LDA.

4.2.3 Predição

As probabilidades marginais para o algoritmo de Gibbs Sampling são dadas por

$$P(a_i = j|\mathbf{a}_{-i}, \mathbf{w}, \varphi) \propto (n^{(d)}_{-i,j} + \alpha_j) \sum_{y=1}^{K} \hat{\varphi}^{(w_i)}_{(j,y)} P(r_{d_i,j} = y|\cdot) \ , \tag{4.7}$$

onde o termo $P(r_{d_i,j} = y|\cdot)$ é igual ao modelo Nota-Aspecto dado pela Equação (4.4). Da mesma forma, as notas podem ser estimadas através da mesma Equação (4.5) do modelo Nota-Aspecto.

5 EXPERIMENTOS

Neste capítulo são apresentados os resultados obtidos em experimentos com os modelos de tópicos descritos em detalhe no Capítulo 3 na representação de resenhas de usuário. A Seção 5.1 descreve o conjunto de dados utilizado nos experimentos realizados. Nas Seções 5.2 e 5.3 são apresentados resultados da estimação dos modelos no conjunto de treinamento, que possibilitam uma análise qualitativa dos tópicos e das representações de documentos em baixa dimensionalidade, respectivamente. A Seção 5.4 apresenta uma análise quantitativa do desempenho dos modelos utilizando-se das representações estimadas para classificar as resenhas em relação a múltiplos aspectos através da aplicação de classificadores estado-da-arte. Por fim, os resultados apresentados são discutidos na Seção 5.5

5.1 Conjunto de dados

Os experimentos apresentados aqui utilizam um conjunto de documentos de resenhas de usuário extraídos da internet a partir do sítio we8there (`http://www.we8there.com`). Foram extraídas 6260 resenhas escritas em língua inglesa contendo comentários sobre restaurantes de ramos diversos (e.g., cozinha italiana, frutos do mar, cafeterias) e geograficamente dispersas (embora a maior parte consista de estabelecimentos nos EUA), de forma a garantir um conjunto bastante heterogêneo e representativo de resenhas de restaurantes. Cada documento consistia originalmente de textos curtos (em média 90 palavras) e um vetor contendo cinco notas numa escala de 1 a 5 avaliando os seguintes aspectos: *comida*, *serviço*, *valor*, *ambiente* e *experiência*. Os dados foram pré-processados, extraindo-se unigramas como atributos através da simples separação de palavras por espaços. Devido à grande quantidade de ruído presente nas resenhas, como erros de ortografia, gírias e abreviações, esta abordagem de segmentação simples monstrou-se mais bem sucedida que métodos mais complexos como *tokenizadores* por expressões regulares ou algoritmos estatísticos. Unigramas com freqüência menor que 10 ou pertencentes a uma lista de palavras

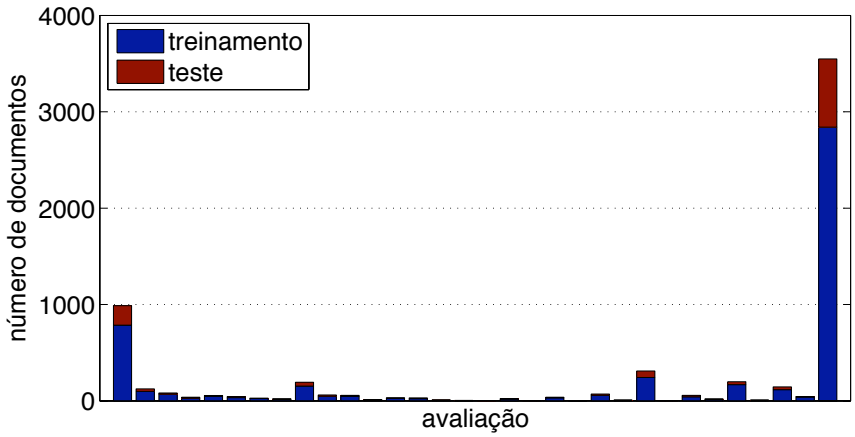


Figura 6: Histograma das diferentes combinações de avaliações presentes no conjunto de dados de we8there.

de função foram removidos. O conjunto final contém 3402 palavras distintas (vocábulos). As notas foram polarizadas, transformando-se avaliações iguais ou inferiores a 3 em sentimentos negativos e avaliações maiores a 3 em sentimentos positivos. Finalmente, dividiu-se o conjunto na proporção 8/2 de forma a obter-se conjuntos de treinamento e teste utilizados, respectivamente, para estimar tópicos e parâmetros de notas.

A Figura 6 exibe o histograma dos vetores de notas $\langle c, s, v, a, e\rangle$ diferenciados em conjunto de treinamento (porção inferior) e teste (porção superior), onde c, s, v, a, e indicam, respectivamente, as avaliações em relação aos aspectos *comida*, *serviço*, *valor*, *ambiente* e *experiência*. Existem no total $2^5 = 32$ combinações diferentes de avaliações. Como notado por Snyder e Barzilay (2007), a ocorrência de concordância entre as avaliações de múltiplos aspectos é abundante nos dados, o que pode ser evidenciado pelas duas barras mais externas desproporcionalmente maiores que as restantes. De fato, os vetores $\langle neg, neg, neg, neg, neg\rangle$ (987 resenhas,) e $\langle pos, pos, pos, pos, pos\rangle$ (3548 resenhas), correspondentes às barras mais externas na Figura, são responsáveis por aproximadamente 73% do total de avaliações. Ademais, a distribuição de dados é tendenciosa em relação a avaliações positivas. É interessante notar que o conjunto de teste apresenta dados na mesma proporção que o conjunto de treinamentos. A Figura 7 apresenta os histogramas em relação a cada aspecto individualmente, com os conjuntos de treinamento e teste discriminados.

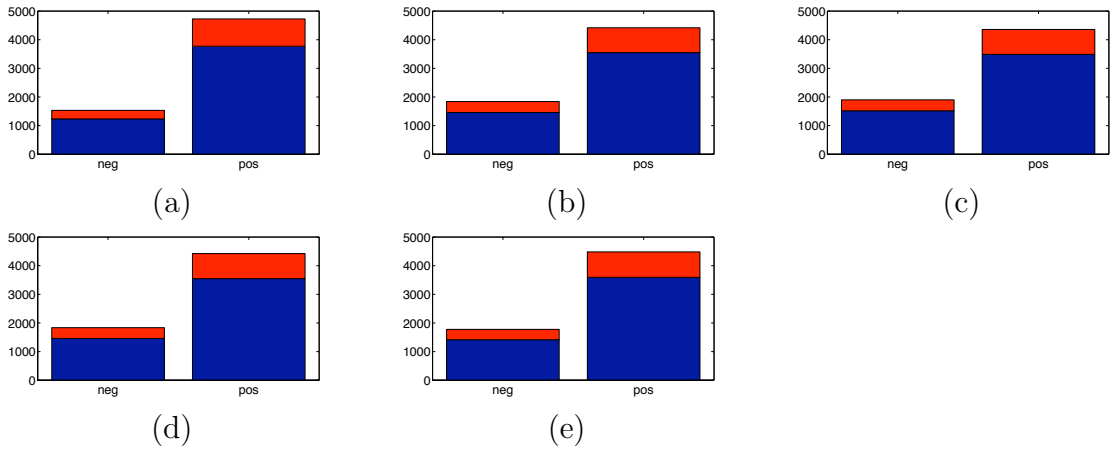


Figura 7: Histogramas das avaliações do conjunto de dados de we8there em relação a um aspecto individual. (a) Comida. (b) Serviço. (c) Valor. (d) Ambiente. (e) Experiência.

5.2 Estimação de tópicos

Os modelos de tópicos representam um documento como uma distribuição sobre tópicos. Dessa forma, uma boa discriminação dos documentos depende, em parte, da qualidade dos tópicos estimados. Nesta seção, são exibidos e discutidos os tópicos estimados pelos modelos unigramas estudados. O modelo de digramas não é analisado aqui devido ao grande número de tópicos gerados.

Existem duas grandes dificuldades à análise qualitativa dos tópicos estimados. A primeira diz respeito à alta dimensionalidade dos vetores que representam os tópicos. O conjunto de dados utilizado nos experimentos contém um vocabulário de 3402 palavras distintas, resultando em tópicos com 3401 graus de liberdade.[1]

A forma usualmente adotada na literatura para exibir tópicos de maneira inteligível é apresentar apenas as palavras mais discriminativas de cada tópico, ou seja, apresentar as palavras que quando observadas mais aumentam a probabilidade de um documento conter um determinado assunto (STEYVERS; GRIFFITHS, 2007; GRIFFITHS; STEYVERS, 2004). Para que isto seja realizado é necessário ordenar as palavras em um tópico segundo uma função de pontuação que afira o poder discriminatório de cada palavra em relação a um tópico.

Blei e Lafferty (2009) utilizam uma função de pontuação para avaliar o k-ésimo vo-

[1] Como os vetores de tópicos devem necessariamente somar uma unidade, uma de suas dimensões é dependente nas restantes.

cábulo do t-ésimo tópico dada por

$$\text{pontuação}(k,t) = \hat{\varphi}_t^{(k)} \log \frac{\hat{\varphi}_t^{(k)}}{\left(\prod_{t=1}^{T} \hat{\varphi}_t^{(k)}\right)^{\frac{1}{T}}} \,. \tag{5.1}$$

Esta função é baseada no esquema de indexação *tf-idf* utilizado em recuperação de informação (MANNING; RAGHAVAN; SCHÜTZE, 2008), e consiste em avaliar o poder discriminatório de uma palavra por duas componentes, uma que avalie a freqüência relativa de uma palavra em um tópico, e outra que avalie a freqüência relativa de uma palavra na coleção (BLEI; LAFFERTY, 2009).

A segunda adversidade à análise qualitativa dos tópicos advém da dificuldade de identificação manual dos tópicos. Estudo empíricos na área de categorização de texto apontam que as funções de decisão estimadas por algoritmos de aprendizagem apresentam termos com alto poder discriminatório que comumente são observados por analisadores humanos como palavras de pouca relevância na determinação da categoria de um documento. Por outro lado, termos intuitivamente tidos como fortemente discriminatórios de uma categoria são, muitas vezes, estimados como de pouca relevância à decisão pela função estimada (SEBASTIANI, 2002).

No estudo realizado por Pang, Lee e Vaithyanathan (2002), onde a habilidade de avaliadores humanos na determinação manual de sentimentos de resenhas de usuário é analisada, constata-se que classificadores baseados em conjuntos de palavras selecionadas manualmente como discriminatórias de um determinado sentimento apresentam baixo desempenho. Por outro lado, palavras estimadas por algoritmos de aprendizagem de máquina como altamente discriminatórias são, muitas vezes, contra-intuitivas. Os autores ilustram o fenômeno com a palavra inglesa "still" que, contra-intuitivamente, apresenta alto poder discriminatório para sentimentos positivos em resenhas de filmes.

A Tabela 2 exibe uma amostra aleatória de dez tópicos estimados com o modelo LDA para $T = 30$, $\alpha = 1,67$ e $\beta = 0,1$, utilizando-se o conjunto de treinamento. Os valores dos hiperparâmetros α e β foram escolhidos com base na literatura (GRIFFITHS; STEYVERS, 2004; STEYVERS; GRIFFITHS, 2007; BLEI; LAFFERTY, 2009). Como apresentado na Seção 5.4 adiante, de maneira geral, os modelos apresentaram melhor desempenho para um número de tópicos próximo a 30. Os tópicos, calculados usando a Equação (3.9) com amostras da milésima iteração do algoritmo Gibbs Sampling, são representados pela dez palavras com maior pontuação de acordo com a Equação (5.1). A maioria dos tópicos parece referir-se a diferente micro-aspectos do aspecto *comida*, como por exemplo os tó-

picos 1, 3 e 7. Os tópicos 5 e 8 contêm palavras comumente utilizadas para descrever o atendimento, podendo serem utilizados para representar o aspecto *serviço*. O tópico 6, interessantemente, parece estar relacionado à opiniões positivas a aspectos diversos, evidenciado pelos adjetivos (e.g., "great", "awesome") e substantivos (e.g., "food", "service"). De maneira geral, e considerando-se as dificuldades discutidas, os tópicos estimados parecem representar os assuntos tratados na coleção utilizada. A tabela completa contendo todos os tópicos encontra-se no Anexo A.

Tabela 2: Amostra dos tópicos estimados com o modelo *LDA* no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem decrescente de pontuação para cada tópico.

Tópico	**10 Palavras mais pontuadas**
0	know make just way want don't sure people going eat
1	breakfast cafe lunch diner coffee day local sunday hotel eggs
2	restaurant family steak special visit wife tea meal feel children
3	served sauce garlic fresh grilled cheese salad mushrooms entrees potatoes
4	best i've try favorite town times area eaten we've far
5	minutes wait table seated arrived hour order waited took waiter
6	great food place service love atmosphere awesome wonderful fantastic fabulous
7	fries sandwich burger cheese menu burgers lunch bacon sandwiches really
8	manager did owner told asked rude service money experience said
9	new years home location stop restaurant old wings ago trip

Na Tabela 3 é exibida uma amostra aleatória de dez tópicos estimados com o modelo de temas para $T = 30$, $\alpha = 1,67$ e $\beta = 0,1$. A numeração dos tópicos é apenas ilustrativa, isto é, não é possível estabelecer relação entre tópicos de amostras diferentes através do índice. Novamente, há uma prevalência de tópicos relacionados ao aspecto *comida* (tópicos 0, 1, 2 e 7) e tópicos relacionados ao aspecto *serviço* (tópico 5) e a boas avaliações (tópico 3). O tópico 2 em particular parece ilustrar más avaliações referentes ao aspecto *comida*. Novamente, a tabela completa contendo todos os tópicos encontra-se no Anexo A. Pelas tabelas apresentadas nesta seção e também pelas exibidas em anexo, não é possível comparar qualitativamente os tópicos estimados pelo modelo LDA e pelo modelos de temas. De fato, ambos os modelos apresentam tópicos bem similares, onde a principal diferença é o ranqueamento das palavras dentro de cada tópico de acordo com a função de pontuação.

As Tabelas 4 e 5 exibem, respectivamente, amostras aleatórias de dez tópicos estimados com os modelo Nota-Aspecto e Nota-Aspecto LDA, para $T = 30$ (5 aspectos observados rotulados em positivos e negativos e 25 aspectos latentes com apenas um rótulo),

Tabela 3: Amostra dos tópicos estimados com o modelo de *temas* no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem decrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	breakfast coffee eggs chocolate cream toast brunch cafe diner ice
1	pizza cheese chicago crust pizzas deli italian pepperoni shop sandwiches
2	ordered salad food bland overcooked good husband steak service tasted
3	great food service friendly excellent wonderful atmosphere staff place recommend
4	years best new restaurant family year live crab stop wish
5	minutes waitress asked table told waiter came seated waited took
6	it's place don't just know i'm want eat like didn't
7	sauce garlic chicken shrimp salad grilled cheese served crab fresh
8	bar great music pub beer irish place vegas beers german
9	wine experience chef list duck presentation bistro amazing course dessert

$\alpha = 1,67$ e $\beta = 0,1$. A coluna mais à esquerda da tabela indica o aspecto e o sentimento a que se refere o tópico representado pelas palavras à direita. Diferentemente dos outros modelos, nos modelos Nota-Aspecto e Nota-Aspecto LDA os tópicos são identificáveis, pois são indexados por um aspecto e sentimento. As tabelas completas podem ser vistas no Anexo A. Como é possível observar na tabela, os tópicos estimados, representados pelas dez palavras de maior pontuação, são coerentes com o par aspecto, sentimento que representam.

Tabela 4: Amostra dos tópicos estimados com o modelo *Nota-Aspecto* no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem decrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	bad cold terrible worst bland told ok dry tasted overcooked
comida+	wonderful great fresh dishes incredible favorite spicy loved grilled italian
serviço-	minutes took waitress finally waited asked brought tip wanted received
serviço+	great friendly staff excellent best world amazing outstanding family fabulous
valor-	did waitress manager money arrived rude wrong ordered paid seated
valor+	love best excellent try fresh reasonable home favorite french homemade
ambiente-	better order rude owner told customers buffet don't tables dirty
ambiente+	great atmosphere wonderful menu delicious dessert nice beautiful unique topped
experiência-	said didn't asked came left manager poor later waitress bad
experiência+	excellent highly great perfect delicious best fantastic enjoy sweet friendly

Os tópicos exibidos nesta seção servem apenas como análise qualitativa primária dos dados estimados pelos modelos de tópicos. Pelos que pode ser observado a partir das tabelas, os tópicos estimados – principalmente nos modelos Nota-Aspecto e Nota-Aspecto

Tabela 5: Amostra dos tópicos estimados com o modelo *Nota-Aspecto LDA* no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem decrescente de pontuação para cada tópico.

Tópico	**10 Palavras mais pontuadas**
comida-	bad ordered terrible worst bland horrible tasted sushi cold ok
comida+	sushi fresh chef delicious dishes incredible roll menu spicy rolls
serviço-	minutes waitress order took waiter waited finally table got brought
serviço+	best excellent family friendly staff italian prices world outstanding home
valor-	didn't better given tables overpriced money think ok charge patrons
valor+	dining french you'll it's walls friday selections south dessert different
ambiente-	owner rude customers card credit kitchen business don't children problem
ambiente+	atmosphere wonderful service excellent delicious beautiful nice cozy great selection
experiência-	said manager asked told did service waitress came tip
experiência+	great love excellent highly recommend best wonderful amazing awesome fantastic

LDA, onde os tópicos são identificáveis – exibem assuntos coerentes ao conjunto de dados utilizado na estimação, de resenhas de restaurantes. Boa parte dos tópicos refere-se a assuntos relacionados ao aspecto *comida*, refletindo a maior proeminência desse aspecto em boa parte das resenhas. A lista completa de tópicos para os modelos "treinados" com 10, 15, 30 e 50 tópicos, representados pelas dez palavras de maior pontuação, encontra-se no Anexo A.

5.3 Estimação de representação de documentos

Nos modelos de tópicos, os documentos são representados por vetores θ cujos elementos indicam o grau de pertinência do documento a um tópico. Se representados adequadamente, documentos que expressam sentimentos similares em relação a aspectos similares devem possuir vetores próximos, enquanto que documentos com sentimentos distintos devem possuir vetores distantes.

Nesta seção, são exibidos os resultados das representações de documentos obtidas pelos modelos de tópicos analisados através da visualização dos vetores θ estimados. Assim como acontece na análise dos tópicos estimados, o elevado número de dimensões dos vetores θ (mesmo que consideravelmente menor que o dos vetores φ) dificulta a análise qualitativa. De forma a superar essa dificuldade, utiliza-se um algoritmo estocástico de projeção bi-dimensional de dados multidimensionais denominada t-SNE (MAATEN, 2008),

que permite a ilustração dos vetores em gráficos planares.[2]

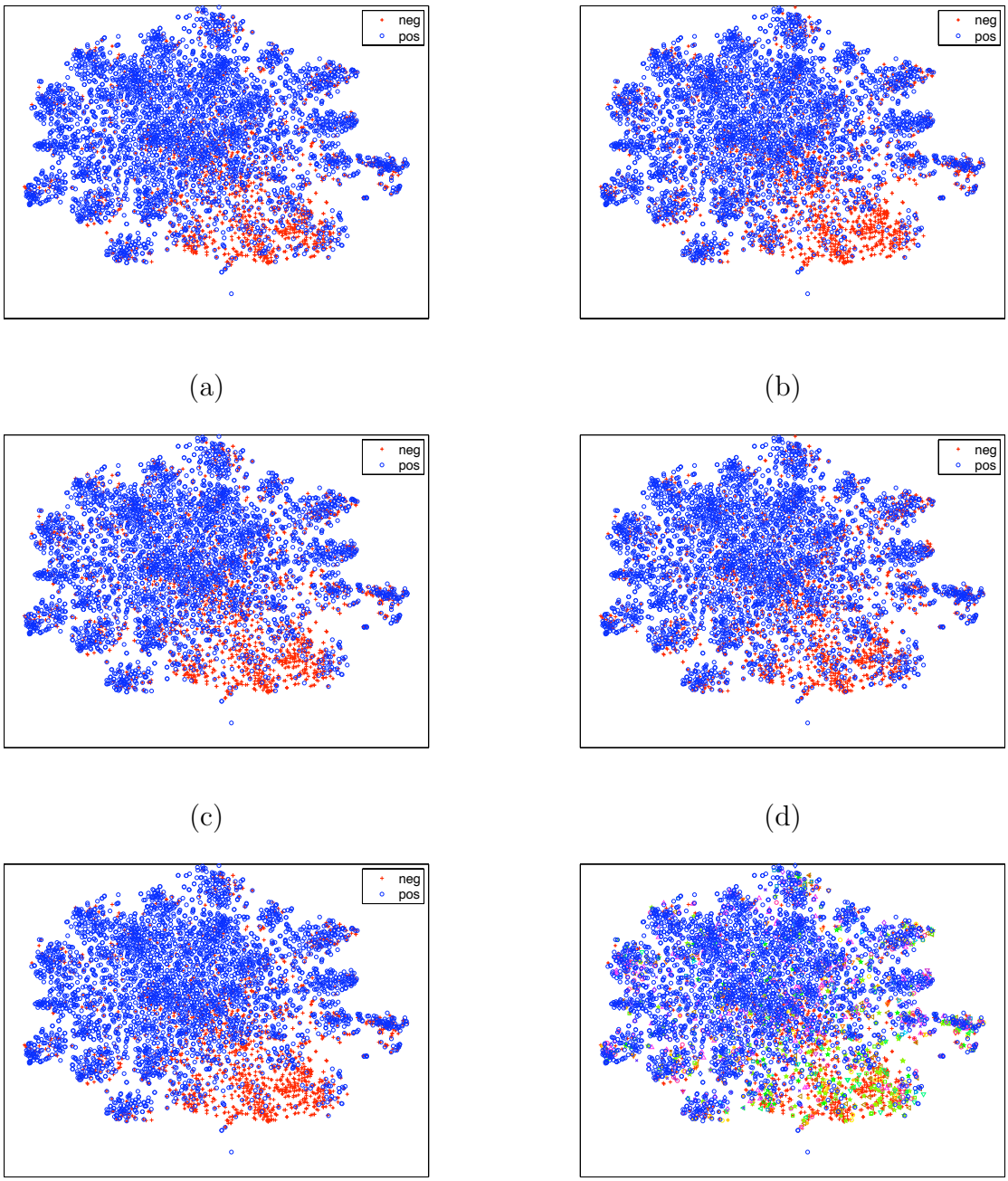
Assim como os demais algoritmos de visualização de dados multidimensionais, o algoritmo t-SNE projeta vetores de uma dimensionalidade maior que dois num espaço de duas dimensões, cuidando para que (i) vetores que em alta dimensão são distantes par-a-par continuem distantes na representação bi-dimensional, e (ii) vetores que em alta dimensão são próximos par-a-par continuem próximos na representação em 2D. Lacoste-Julien, Sha e Jordan (2008) relatam melhores resultados para visualização de seus modelos LDA estendidos utilizando o algoritmo t-SNE.

Seja p uma distribuição sobre distâncias entre os dados na dimensão original. O algoritmo t-SNE realiza uma redução dimensional encontrando a distribuição q bi-dimensional sobre distâncias entre os dados que minimiza a divergência Kullback-Leibler (HASTIE; TIBSHIRANI; FRIEDMAN, 2001) entre as distribuições de alta e baixa dimensionalidade, $KL(p,q) = \sum_x p(x) \log p(x)/q(x)$, onde a somatória faz-se sobre todos os pontos do conjunto de dados (MAATEN, 2008).

As Figuras 8(a)–(e) exibem as projeções em 2D feitas pelo algoritmo t-SNE para os vetores θ estimados pelo modelo LDA a partir do conjunto de treinamento e rotulados em relação a cada um dos aspectos avaliados. Nas figuras, pontos representados por círculos e cruzes indicam, respectivamente, avaliações positivas e negativas em relação ao aspecto avaliado. A Figura 8(f) apresenta a projeção para o vetor conjunto de notas $\langle c, s, v, a, e \rangle$. Círculos e cruzes representam, respectivamente, documentos avaliados positivamente e negativamente em relação a todos os aspectos. Os demais sinais representam avaliações intermediárias. Os vetores de documentos para o modelo LDA foram obtidos a partir da Equação (3.10) utilizando-se amostras da milésima iteração do algoritmo Gibbs Sampling com hiperparâmetros $T = 30$, $\alpha = 1,67$ e $\beta = 0,1$, escolhidos pelos mesmos motivos explicados na Seção 5.2.

É possível notar, comparando-se as projeções obtidas, a ausência de agrupamentos distintos entre avaliações positivas e negativas. De maneira geral, as avaliações negativas concentram-se nos cantos inferiores direito, enquanto que as avaliações positivas ocupam o restante do espaço. Entre as avaliações positivas, é possível notar alguns agrupamentos de dados nas regiões mais externas. As projeções rotuladas em relação aos aspectos *serviço* e *experiência* apresentam menor sobreposição entre resenhas positivas e negativas, se comparadas às demais.

[2] A implementação do algoritmo utilizada neste trabalho pode ser obtida gratuitamente em `http://ticc.uvt.nl/~lvdrmaaten/tsne`.



(a) (b)

(c) (d)

(e) (f)

Figura 8: Representações bi-dimensional obtidas através do algoritmo t-SNE para os vetores de proporção de tópicos θ estimados pelo modelo LDA no conjunto we8there rotuladas em relação a diferentes aspectos . (a) Comida. (b) Serviço. (c) Valor. (d) Ambiente. (e) Experiência. (f) Todos os aspectos.

Vale notar que as projeções bi-dimensionais estimadas pelo algoritmo t-SNE mantêm apenas as características relativas dos dados, ou seja, a visualização é coerente apenas na distância entre pares de pontos; a localização dos pontos não possui qualquer relevância nessas figuras.

As Figuras 9(a)–(d) comparam as representações de documentos dos modelos LDA, LDA ótimo, LDA digrama, de temas, Nota-Aspecto e Nota-Aspecto LDA, respectivamente, projetadas em 2D. Os modelos LDA, digramas, Nota-Aspecto e Nota-Aspecto LDA foram estimados utilizando-se os mesmos valores arbitrários de hiperparâmetros da Figura 8. O modelo LDA ótimo utiliza as Equações (3.12) e (3.13) para estimar os valores do hiperparâmetros α e β, respectivamente. De maneira similar, o modelo de temas utiliza as Equações (3.20) e (3.13) para estimar os valores de α e β, respectivamente.

As figuras referentes aos modelos de digramas, de temas e Nota-Aspecto LDA utilizam os vetores θ estimados de forma semelhante ao modelo LDA. Por não possuir uma representação explícita dos documentos, os vetores de representação de documentos para o modelo Nota-Aspecto foram obtidos através da normalização das freqüências relativas dos rótulos amostrados para cada documento. Nas Figuras, pontos representados por círculos e cruzes indicam, respectivamente, avaliações positivas e negativas em relação ao aspecto *comida*. As representações referentes aos demais aspectos não apresentam diferenças qualitativas significativas às do aspecto apresentado.

Como pode-se notar nas Figuras 9(a) e (b), a estimação dos hiperparâmetros não resulta em ganhos qualitativos na representação de documentos, considerando-se a discriminação entre conjuntos de documentos rotulados como negativos e positivos. Na figura, observa-se que o modelo LDA ótimo apresenta maior coesão entre os agrupamentos intra-classes, isto, é, agrupamentos de dados rotulados com sentimentos semelhantes.

Na representação estimada pelo modelo de digramas, exibida na Figura 9(c), os documentos de classes distintas apresentam-se completamente sobrepostos, demonstrando a menor qualidade do modelo para a representação do sentimento de resenhas. Uma possível explicação para a deterioração da qualidade da representação obtida pelo modelo de digramas, que inclui informação posicional, poder ser dada pelo drástico aumento no número de tópicos (e.g., de 30 para $30 \times 3402 = 102\,060$ nos dados apresentados), resultando num aumento expressivo da dimensionalidade dos vetores de distribuição de tópicos.

A representação obtida pelo modelo de temas na Figura 9(d) apresenta a menor sobreposição de documentos de classes distintas, indicando a maior qualidade do modelo. Assim como no modelo LDA ótimo, existe agrupamentos intra-classes mais delineados.

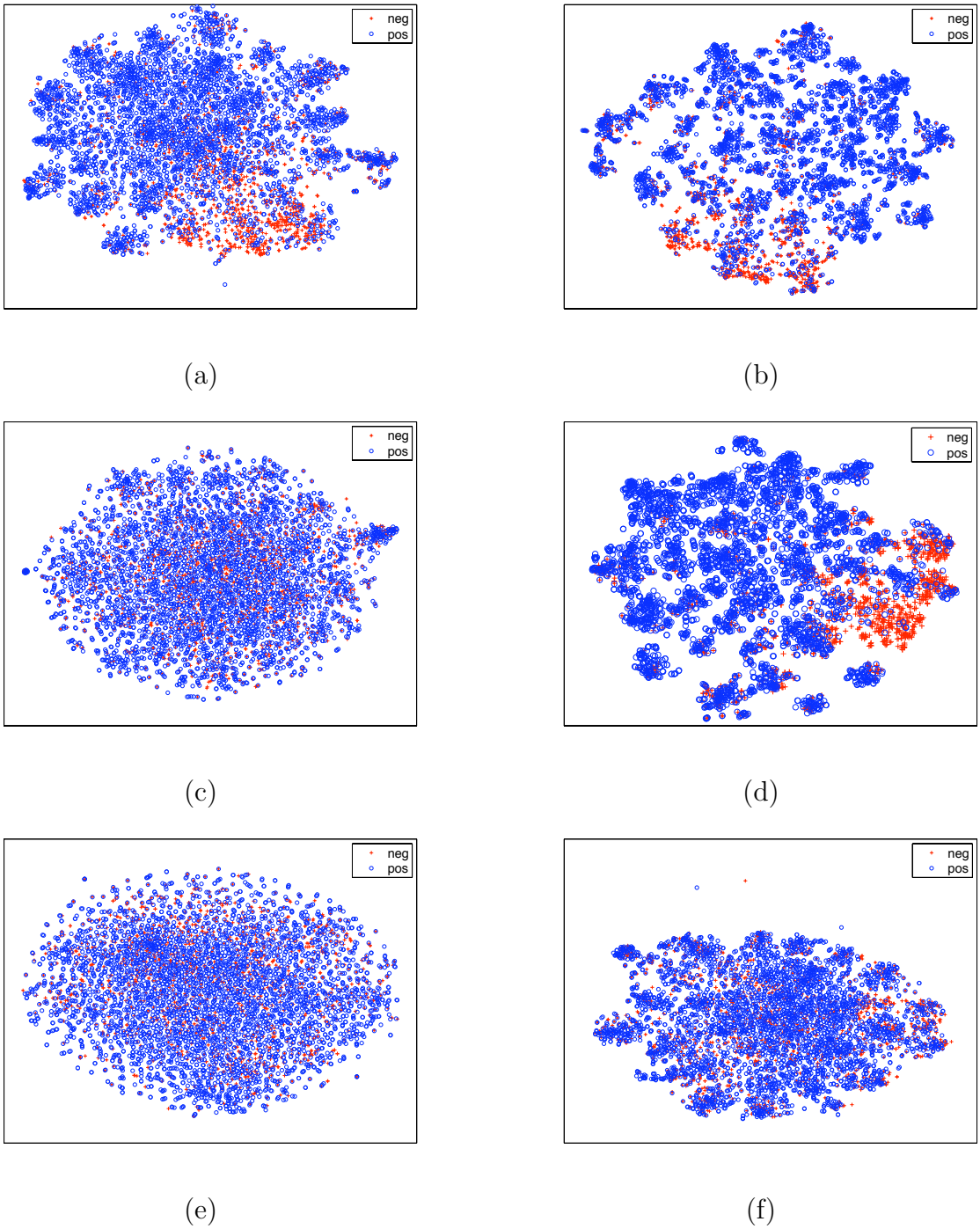


Figura 9: Representações bi-dimensionais obtidas através do algoritmo t-SNE para os vetores de proporção de tópicos θ estimados pelos modelos rotuladas em relação ao aspecto *comida* . (a) LDA. (b) LDA ótimo. (c) Digramas. (d) Temas. (e) Nota-Aspecto. (f) Nota-Aspecto LDA.

Assim como na representação do modelo de digramas, as representações estimadas pelos modelos Nota-Aspecto e Nota-Aspecto LDA (Figuras 9(e) e (f), respectivamente), apresentam uma enorme sobreposição de documentos de classe distintas. Apesar dos tópicos serem coerentemente estimados, esses modelos falham em representar os documentos pelos vetores de proporções de tópicos.

Pelas visualizações apresentadas nas figuras, o modelo de temas exibe melhor qualidade de representação de documentos, seguido pelo modelo LDA. Os modelos de digramas e que incluem informações de Nota e Aspecto mostraram-se incapazes de discriminar os documentos em relação aos sentimentos expressos. Contrariando o trabalho recente de Asuncion et al. (2009), a estimação dos hiperparâmetros α e β não resultou em melhora significativa do modelo estimado. O modelo de temas é um caso particular, pois sua diferença em relação ao modelo LDA original consiste na estimação dos hiperparâmetros α_ℓ dependentes nas categorias dos documentos. Por fim, cabe ressaltar que as projeções apresentadas são obtidas por estimação a partir dos dados multidimensionais, estando, portanto, sujeitas a erros introduzidos no processo de redução de dimensionalidade realizado pelo algoritmo t-SNE.

5.4 Classificação de sentimentos

Os resultados qualitativos apresentados nas seções anteriores são interessantes para ter-se uma idéia da qualidade das diferentes componentes dos modelos estimados a partir dos dados. Entretanto, o objetivo final deste trabalho é encontrar uma representação que, de fato, ajude a aumentar o desempenho dos classificadores de texto baseados em aprendizagem de máquina na tarefa de discriminação de sentimentos de resenhas de usuários.

Esta seção apresenta resultados que permitem uma análise quantitativa da qualidade das representações obtidas pelas diversas configurações através da acurácia obtida na classificação de documentos em relação aos múltiplos aspectos avaliados pelos usuários. São utilizados dois classificadores estado-da-arte na tarefa de categorização de texto, Máxima Entropia (MaxEnt) e Support Vector Machine (SVM), descritos no Capítulo 2.[3]

Na Subseção 5.4.1 são analisadas diferentes configurações de atributos utilizando-se as representações de tópicos fornecidas pelos modelos de tópicos. Os resultados são comparados à representação *bag of words* usual. Na Subseção 5.4.2, o efeito da estimação

[3]As implementações para os classificadores utilizadas neste trabalho podem ser encontradas em `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html` e `http://svmlight.joachims.org/`.

dos hiperparâmetros α e β frente ao objetivo final é analisado. Por fim, as Subseções 5.4.3 e 5.4.4 avaliam a influência da incorporação de informações posicionais e categóricas, respectivamente, no modelo LDA original.

5.4.1 Extração de atributos

A representação mais comumente adota na representação de documentos de texto, denominada *bag of words*, consiste em representar documentos por vetores cujos elementos indicam a freqüência de ocorrência de cada palavra do vocabulário no texto. Esse modelo de representação é denotado aqui modelo de *unigramas*. Os modelos de tópicos fornecem uma representação alternativa substituindo a freqüência de ocorrência de palavras pela freqüência de ocorrência de tópicos no texto. Essa representação é denotada aqui por representação de *tópicos*. Os resultados apresentados nas seções anteriores focaram na análise qualitativa da representação de tópicos fornecida pelos diferentes modelos. Entretanto, do ponto de vista do problema de classificação, é possível combinar os atributos obtidos nesses dois modelos de representações distintos (unigramas e tópicos) formando-se novas representações de documentos. Neste trabalho, duas representações que combinam atributos de tópicos e unigramas são estudadas.

Os modelos de tópicos representam cada posição de um documento de texto por um par $\langle w_i, z_i \rangle$, onde w_i indica o unigrama da i-ésima posição e z_i indica seu rótulo na representação de tópicos. Assim, as representações de unigramas e tópicos particionam de fato esta representação utilizando apenas uma parte dos dados (unigramas ou rótulos).

A representação denominada *tópicos*unigramas*, inspirada na representação utilizada em Mauá e Cozman (2009) e ilustrada na Figura **??**, consiste em representar um documento pela freqüência de pares de unigrama, tópico $\langle w_i, z_i \rangle$. Dessa forma, é possível representar a dependência entre as palavras e ainda manter o poder discriminatório de cada vocábulo. Os vetores de representação de documentos, no entanto, apresentam um ganho de dimensionalidade e são representados utilizando-se $T * W - 1$ graus de liberdade (contra $T - 1$ e $W - 1$ graus de liberdade para as representações de tópicos e unigramas, respectivamente).

A segunda representação, denominada *tópicos e unigramas*, consiste em formar vetores através da simples concatenação dos vetores unigramas e tópicos. Seja u um vetor representando as freqüências de unigramas e v um vetor representando as freqüências de tópicos em um documento, o vetor $w = \langle u, v \rangle$ representa o documento no modelo de *tópicos e unigramas*.

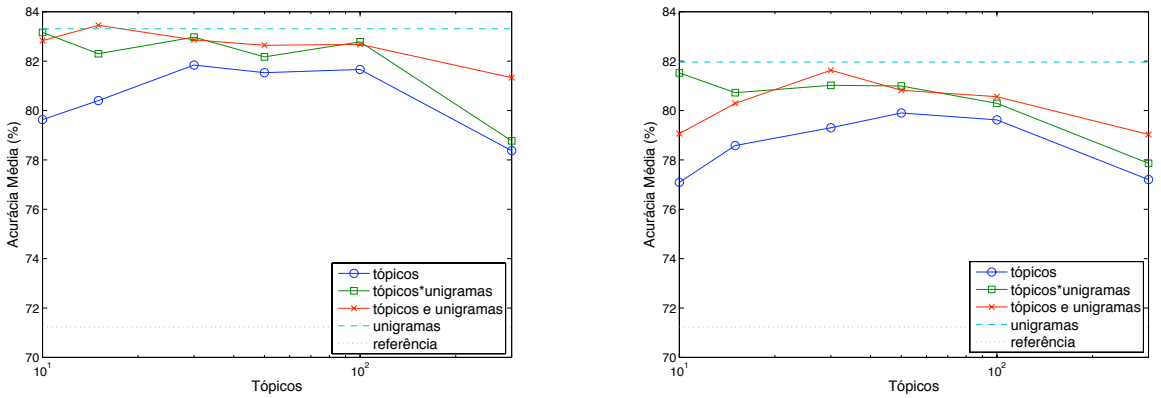


Figura 10: Acurácia média da tarefa de predição de avaliações multiaspectos pelo número de tópicos para diferentes representações utilizando rótulos extraídos do modelo LDA no conjunto de dados de we8there. (a) Máxima Entropia. (b) SVM.

A Figura 10 apresenta a acurácia média de dois classificadores (MaxEnt e SVM) na tarefa de classificação de sentimentos baseada em múltiplos aspectos em função do número de tópicos, utilizando atributos estimados pelo modelo LDA ($\alpha = 50/T$, $\beta = 0.1$). A acurácia média é dada pela média das acurácias das tarefas de classificação de sentimento de resenhas em positivo ou negativo em relação a cada um dos aspectos avaliados no conjunto de dados de teste utilizado. Os rótulos z_i utilizados nas representações foram obtidos pela moda da distribuição de amostras realizadas pelo algoritmo Gibbs Sampling com 1000 iterações e *burn-in* de 500 iterações. A linha horizontal inferior em cada um dos gráficos ilustra o desempenho de um classificador de referência que decide sempre pelo sentimento mais abundante nos dados em relação a um aspecto; particularmente para o conjunto de dados utilizado o classificador de referência decide sempre positivo.

Pela figura, é possível constatar que o classificador de Máxima Entropia apresenta desempenho superior ao classificador SVM. A representação de tópicos apresenta desempenho bastante inferior à representação de unigramas, obtendo melhores resultados para um número de tópicos menor que 50. A representação de tópicos*unigramas exibe desempenho próximo à de unigramas para um número relativamente pequeno de tópicos, e desempenho próximo à de tópicos para um número grande de tópicos. De forma similar, a representação de tópicos e unigramas desempenha ligeiramente melhor que à unigramas no classificador MaxEnt quando $T = 15$, decaindo com o aumento do número de tópicos. Para o classificador SVM, o melhor desempenho obtido em $T = 30$ ficou abaixo do desempenho da representação de unigramas. Por fim, todas os modelos apresentaram desempenho bastante superior ao classificador de referência.

É interessante ressaltar o fenômeno de *maldição da dimensionalidade* que faz com que a acurácia de algoritmos de aprendizagem de máquina caia drasticamente quando o número de dimensões dos dados é muito alto, em geral quando o número de dimensões é da ordem do número de instâncias utilizadas no treinamento (HASTIE; TIBSHIRANI; FRIEDMAN, 2001; SEBASTIANI, 2002). Ambas as representações de tópicos*unigramas e tópicos e unigramas geram dados com número maior de dimensões (34020–1020600 e 3412–3702, respectivamente) que o número de documentos utilizado no treinamento (5008).

De maneira geral, a inclusão de atributos estimados pelo modelo LDA não resulta em aumento de desempenho na determinação de sentimentos. De fato, corroborando o resultado qualitativo demonstrado na Figura 8, a representação de tópicos obtida pelo modelo LDA não ajuda na discriminação de documentos de sentimentos diferentes. Entretanto, cabe ressaltar que a representação de tópicos realiza uma redução drástica na dimensionalidade dos dados, por exemplo das 3402 dimensões do modelo de unigramas para 30 dimensões no modelo LDA com 30 tópicos.

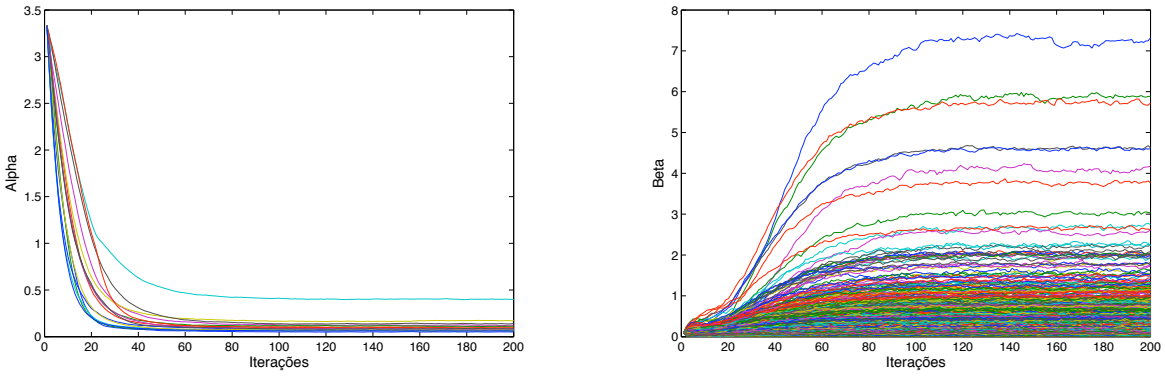
5.4.2 Estimação de hiperparâmetros

Na literatura de modelos de tópicos, é comum o uso de valores arbitrários para os hiperparâmetros α e β (BLEI; LAFFERTY, 2009; GRIFFITHS et al., 2005; GRIFFITHS; STEYVERS, 2004). Asuncion et al. (2009), entretanto, alerta para a importância da estimação dos hiperparâmetros no desempenho do modelo estimado.

Esta subseção investiga a relevância da estimação dos hiperparâmetros na classificação de sentimentos. Os parâmetros são estimados utilizando-se o método de ponto fixo do Algoritmo 2.

Os valores dos vetores α e β para o modelo LDA com 15 tópicos são apresentados em função do número de iterações na Figura 11. Os valores iniciais dos parâmetros são os valores arbitrários comumente utilizados na literatura, $\alpha = 50/T = 3,33$ e $\beta = 0,1$. Os valores de α convergem após pouco mais de 80 iterações a valores próximo a $0,1$, com exceção de um elemento, relativo ao tópico 3 da Tabela 11, que converge para um valor próximo a $0,4$. Os valores de β, por sua vez, convergem após 100 iterações em valores no intervalo $0,1$–$7,2$.

Para entender o significado desses valores, lembre-se do processo generativo do modelo LDA, onde os parâmetros α e β são utilizados, respectivamente, para amostrar os vetores



(a) (b)

Figura 11: Valores estimados dos hiperparâmetros α_j e β_k pelo algoritmo de ponto fixo em função do número de iterações do algoritmo EM. (a) α. (b) β.

de documento θ e de tópicos φ a partir de distribuições Dirichlet.

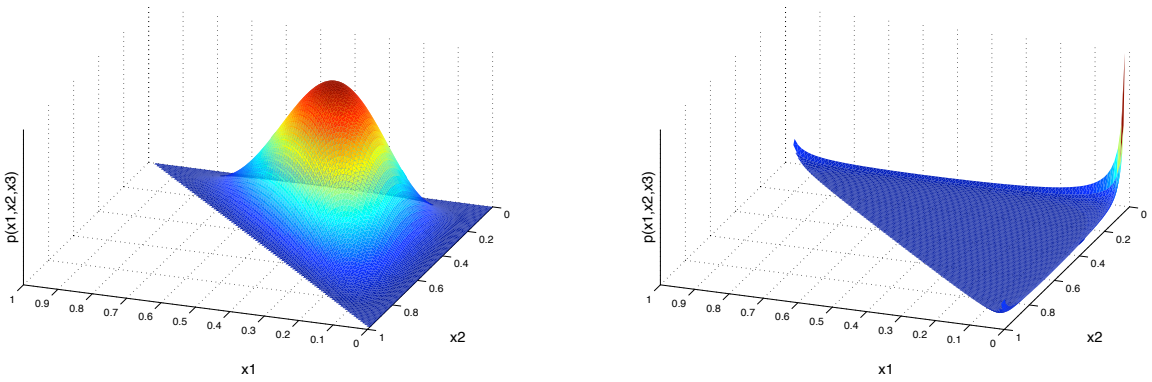
A distribuição Dirichlet é uma distribuição contínua sobre pontos no K-simplex, isto é, sobre vetores $\mathbf{x} = \langle x_1, \ldots, x_K \rangle$, onde $x_i \geq 0$ e $\sum_{i=1}^{K} x_i = 1$ (MINKA, 2003). A função de distribuição é dada por

$$\text{Dirichlet}(\mathbf{x}; \alpha) = p(\mathbf{x}) = \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \ ,$$

onde $\alpha = \langle \alpha_1, \ldots, \alpha_K \rangle$, $\alpha_i > 0$.

A forma da distribuição é bastante flexível e dependente nos parâmetros α. Para $\alpha = 1$, a distribuição torna-se uniforme. Para $0 < \alpha < 1$, a distribuição concentra-se nos vértices do simplex, aumentando a heterogeneidade dos pontos amostrados. Fenômeno inverso ocorre quando $\alpha > 1$, onde a distribuição toma forma de "sino", concentrando sua massa no centro do simplex, o que produz pontos amostrados próximos.

Como ilustração, tome o caso da distribuição *a priori* de vetores de documentos θ para um modelo LDA de 3 tópicos. Para o valor de α comumente utilizado na literatura (i.e., α simétrico com elementos $\alpha_i = 3,33$), tem-se a distribuição ilustrada na Figura 12(a). Pela figura, é possível notar que a maior parte da massa da distribuição concentra-se em torno do ponto $\mathbf{x} = \langle 0,33\ 0,33\ 0,33 \rangle$, aumentando a probabilidade de documentos com misturas homogêneas de tópicos (i.e., proporções similares entre todos os tópicos). Quando utilizada como distribuição *a priori*, essa distribuição Dirichlet tem o efeito de "suavizar" as distribuições empíricas θ estimadas a partir dos vetores $\mathbf{z}$ amostrados, reforçando a homogeneidade da proporção entre tópicos (vide Equação (3.10)). Este efeito parece razoável levando-se em conta que resenhas de usuários em geral contêm misturas homogêneas de



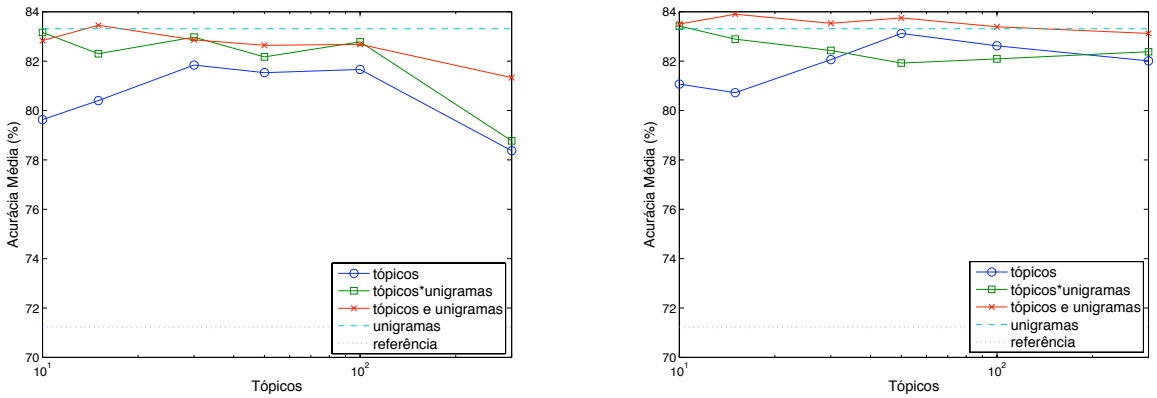
(a) (b)

Figura 12: Exemplos de distribuições Dirichlet com parâmetros distintos. (a) $\alpha = \langle 3,33\ 3,33\ 3,33 \rangle$. (b) $\alpha = \langle 0,1\ 0,1\ 0,4 \rangle$.

tópicos refletindo as avaliações dos múltiplos aspectos. Assim, caso os tópicos estimados refiram-se aos aspectos opinados, é razoável esperar vetores θ concentrados em torno de um ponto central.

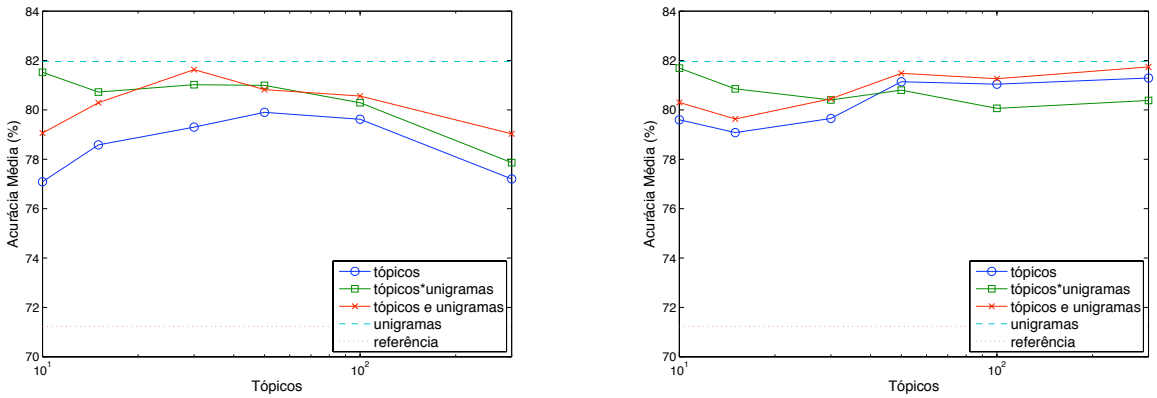
Diferentemente do esperado, o parâmetro α estimado pelo algoritmo de estimação de ponto fixo apresenta valores de $\alpha < 1$, com um valor de α_3 maior que a média. Esse valor indica que, de fato, os vetores são heterogeneamente distribuídos e contêm apenas poucos tópicos proeminentes, sendo o tópico 3 o mais abundante. A Figura 12(b) ilustra uma distribuição Dirichlet para um vetor de parâmetros $\alpha = \langle 0,1\ 0,1\ 0,4 \rangle$ semelhante ao estimado (considerando-se a diferença de dimensionalidade entre os casos). Pode-se notar que a distribuição concentra-se nos vértices, em particular no vértice cujo parâmetro respectivo tem maior valor. Essa distribuição afere probabilidades pequenas para documentos que contenham tópicos em proporções semelhantes, e privilegia documentos contendo um dos tópicos majoritariamente.

Pela Tabela 11, que ilustra os tópicos estimados para o modelo LDA com 15 tópicos, observa-se a abundância de tópicos relacionados a diferentes facetas do aspecto *comida*. Note também que o tópico 3, cujo parâmetro α_i estimado possui maior valor, contém palavras utilizadas freqüentemente na avaliação de qualquer aspecto, além de palavras relacionadas aos aspectos. Essas evidências indicam que, contrário ao esperado, o modelo LDA captura tópicos relacionados às diferentes classes de restaurantes encontrados no conjunto de dados (e.g., frutos do mar, comida italiana, comida indiana), produzindo principalmente representações para documentos baseadas nas diferentes particularidades do objeto avaliado, ao invés de produzir representações que capturem os sentimentos expressos.



(a) (b)

Figura 13: Acurácia média do classificador MaxEnt na classificação de sentimento para diferentes parâmetros α, β e T para o modelo LDA no conjunto de dados de we8there. (a) $\alpha = 50/T$, $\beta = 0.1$. (b) α e β estimados.



(a) (b)

Figura 14: Acurácia média do classificador SVM na classificação de sentimento para diferentes parâmetros α, β e T para o modelo LDA no conjunto de dados de we8there. (a) $\alpha = 50/T$, $\beta = 0.1$. (b) α e β estimados.

As Figuras 13 e 14, exibem a acurácia média dos classificadores MaxEnt e SVM, respectivamente, na classificação de sentimentos. As Figuras 13(a) e 14(a) exibem o desempenho para representações obtidas através de modelos LDA estimados com valores arbitrários de α e β normalmente utilizados. Nas Figuras 13(b) e 14(b), são exibidos os resultados para representações obtidas de modelos LDA com os valores dos hiperparâmetros estimados pelo método de ponto fixo.

Com exceção da representação tópicos*unigramas, o modelo com parâmetros estimados obteve desempenho superior ao modelo com parâmetros arbitrários. Em particular, a representação de tópicos e unigramas apresenta desempenho superior ao modelo *bag of words* em relação ao classificador MaxEnt. No entanto, para um número de tópicos me-

nor que 50, o desempenho das representações, apesar de superior, não apresenta diferença significativa, principalmente se considerando o enorme custo computacional da estimação dos hiperparâmetros.

5.4.3 Incorporando informação posicional

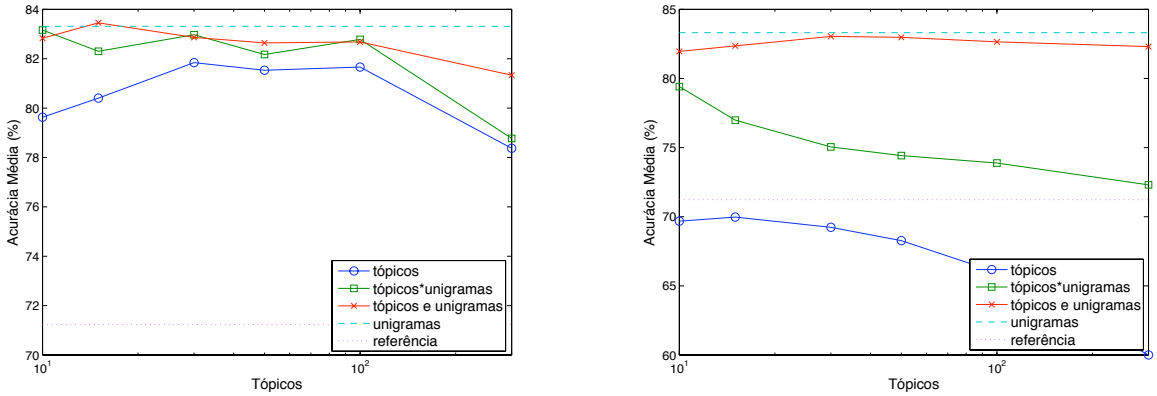
Os modelos de tópicos, em geral, ignoram a ordem das palavras no texto, fazendo com que as informações de relações entre as palavras sejam perdidas. O modelo de digramas recupera em parte essas informações, incluindo digramas, pares de palavras, no modelamento de documentos. A inclusão de digramas permite a inclusão de informação contextual, porém aumenta drasticamente a dimensionalidade do modelo (WALLACH, 2006). Nesta seção, analisa-se os efeitos da incorporação dessa informação nas representações obtidas a partir dos rótulos estimados.

As Figuras 15 e 16 exibem os gráficos da acurácia média dos obtidas com representações dos dois modelos utilizando-se classificadores MaxEnt e SVM, respectivamente. Como é possível observar, o desempenho da classificação de sentimentos é imensamente deteriorado quando são incluídos atributos dos modelos de digramas. No classificador MaxEnt, a representação de tópicos obtém desempenho pior que o classificador de referência. O único desempenho parcialmente satisfatório obtido com rótulos do modelo de digramas é observado na representação de tópicos e unigramas do gráfico do classificador MaxEnt (Figura 15(b)). No entanto, devido à natureza deste classificador, é possível que este desempenho resulte da utilização apenas dos atributos de unigramas, atribuindo pesos ínfimos aos atributos de digramas. Esses resultados confirmam os resultados da análise qualitativa das representações de documento na Seção 5.2.

5.4.4 Incorporando informações categóricas

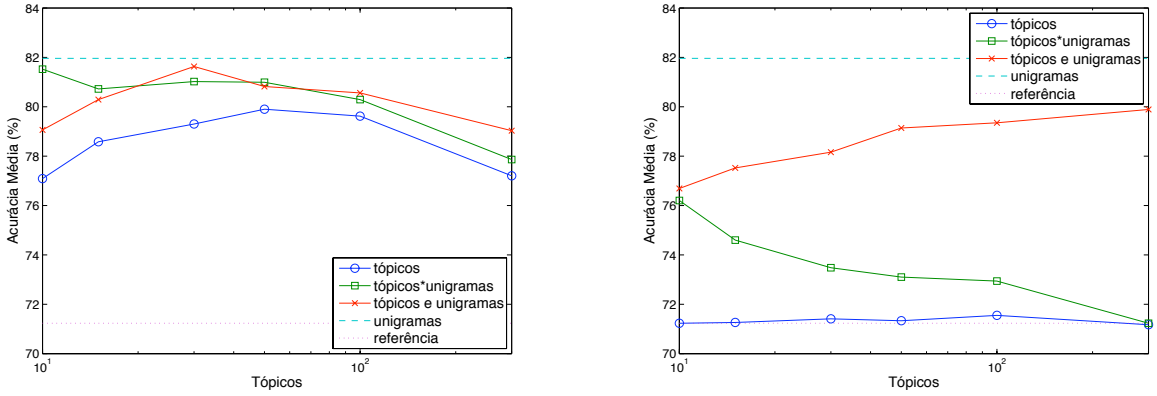
O modelo LDA não permite a inclusão da informação das avaliações e aspectos presente no conjunto de dados na estimação dos dados (treinamento). Esta seção analisa a inclusão desta informação no desempenho da tarefa de classificação de sentimentos.

Modelo de temas O modelo de temas incorpora informação de avaliação (sentimento) sobre notas através da utilização de distribuições *a priori* distintas para cada avaliação (sentimento). Entretanto, o modelo permite que cada documento contenha apenas uma avaliação. De forma a permitir a incorporação das informações referentes aos múltiplos



(a) (b)

Figura 15: Acurácia média do classificador MaxEnt para as versões unigrama e digrama do modelo LDA no conjunto de dados de we8there. (a) LDA. (b) LDA Digrama.



(a) (b)

Figura 16: Acurácia média do classificador SVM para as versões unigrama e digrama do modelo LDA no conjunto de dados de we8there. (a) LDA. (b) LDA Digrama.

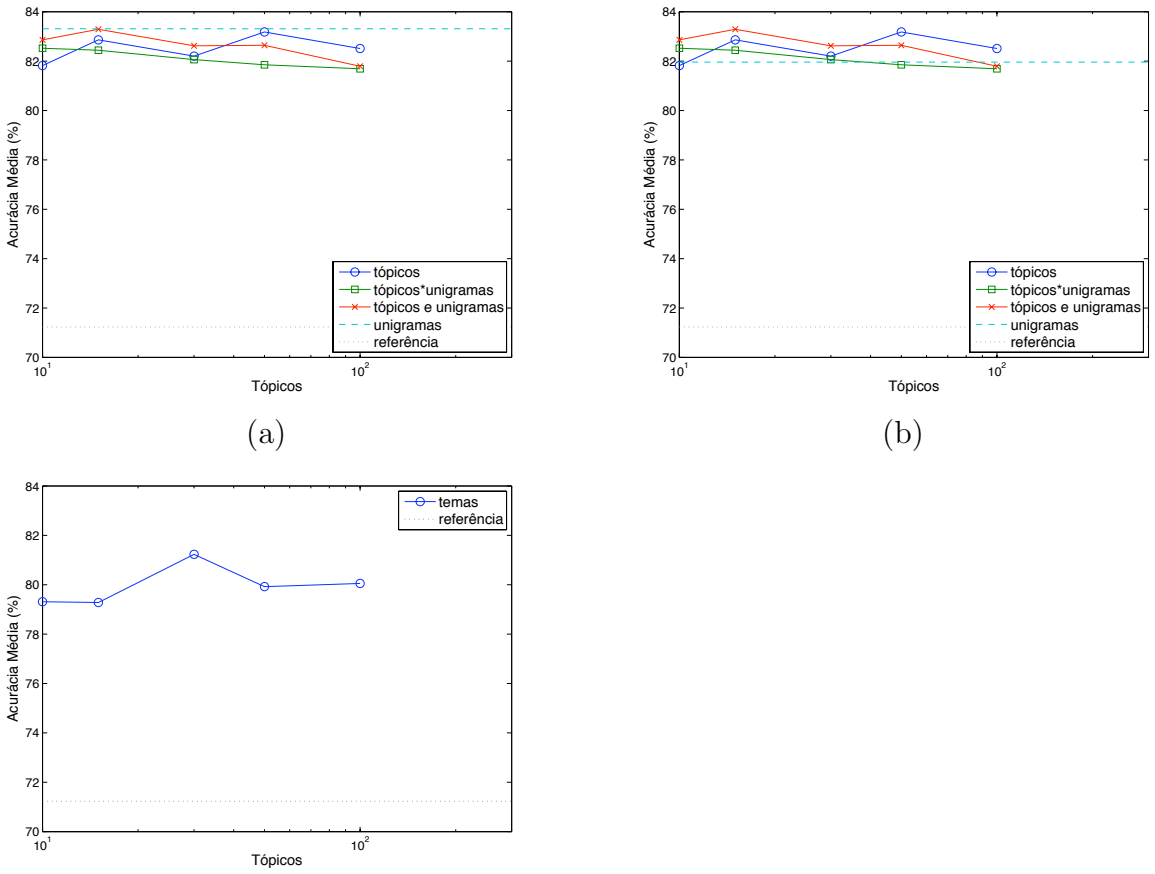


Figura 17: Acurácia média do modelo de temas utilizando diferentes classificadores. (a) Máxima Entropia. (b) SVM. (c) Modelo de Temas

aspectos opinados, treinou-se um modelo de tema para cada aspecto do conjunto de aspectos avaliados utilizando-se apenas as avaliações referentes ao aspecto. Os rótulos foram então utilizados nas diversas representações para classificar os documentos em relação ao aspecto utilizado na estimação.

A Figura 17 apresenta os resultados utilizando dados estimados pelos modelos de temas. Além dos classificadores MaxEnt e SVM, o modelo de temas permite a inferência automática das avaliações de sentimento, cujo desempenho é ilustrado na Figura 17.

Como ressaltando anteriormente, o modelo de temas incorpora a informação de categorias através da estimação de hiperparâmetros distintos para cada categoria. Esse procedimento custoso é então repetido para cada aspecto, o que leva estimações com custos computacionais proibitivos. Por essa razão os modelos de temas foram estimados com um número máximo de 100 tópicos. A estimação para modelos com 300 tópicos levaria mais de um mês de processamento ininterrupto.

Os resultados demonstram que, de maneira geral e a exemplo da análise qualitativa da estimação dos documentos, o modelo de temas apresenta desempenho ligeiramente superior ao modelo LDA original. Este aumento de desempenho evidencia, mais uma vez, a influência das distribuições *a priori* na qualidade da estimação do modelo. No entanto, tal aumento vem acompanhado de um enorme custo computacional. Apesar de o modelo de temas apresentar bons desempenhos nas diversas representações e configurações de número de tópicos, o melhor resultado obtido ($83,29\%$ na representação de tópicos e unigramas) ficou abaixo do melhor resultado obtido com o modelo original ($83,90\%$ na mesma representação).

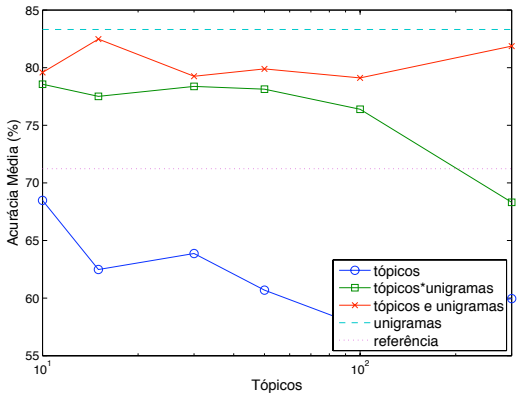
Por fim, cabe notar que a utilização das avaliações inferidas pelo modelo a partir das amostras como método de classificação obteve um mau desempenho.

Modelo Nota-Aspecto Diferentemente do desejado, os tópicos estimados pelo modelo LDA relacionam-se às diferentes subcategorias presentes nos dados, e apenas em menor parte aos aspectos avaliados. Os modelos Nota-Aspecto e Nota-Aspecto LDA tentam solucionar o problema condicionando os tópicos estimados às informações de avaliações multiaspectos presente nos dados.

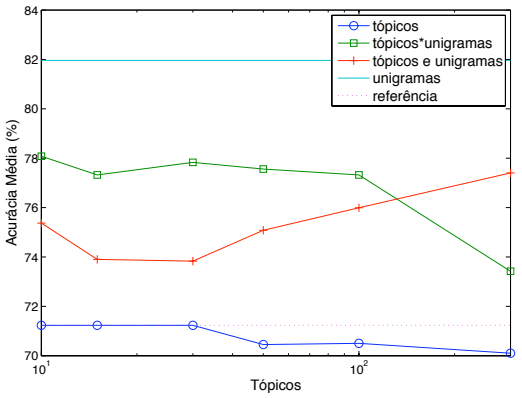
Como ilustrado na Seção 5.2, os tópicos estimados pelos modelos Nota-Aspecto e Nota-Aspecto LDA distinguem as múltiplas opiniões contidas nos textos e endereçadas aos diferentes aspectos. Existe, no entanto, um número maior de aspectos comentados no texto que o número de aspectos rotulados encontrado no conjunto de dados. De forma a permitir maior flexibilidade ao modelo incluindo essas opiniões os modelos foram estimados assumindo que os tópicos referentes a aspectos cujas avaliações não estão presentes nos dados referem-se a atributos que possuem apenas uma avaliação possível. Este procedimento foi adotado nos modelos Nota-Aspecto e Nota-Aspecto LDA com mais de 10 tópicos.

A Figura 18 exibe o desempenho das representações geradas pelo modelo Nota-Aspecto para os classificadores MaxEnt , SVM e Nota-Aspecto. Repetindo os resultados obtidos na análise qualitativa da representação de documentos, os atributos estimados levam a péssimos resultados na determinação de sentimentos da resenhas.

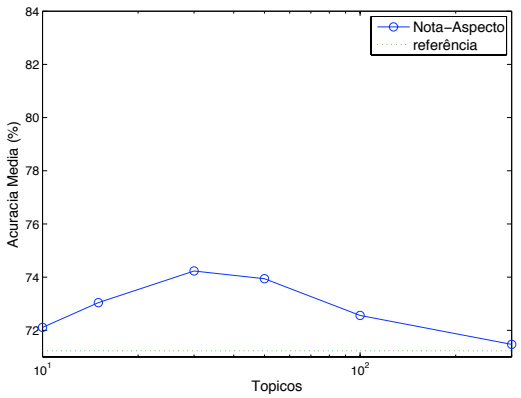
Apesar dos tópicos estimados pelo modelo parecerem descrever bem as opiniões contidas no texto, o modelo Nota-Aspecto gera representações insatisfatórias para os documentos. Na Figura 21 são exibidas algumas amostras de resenhas rotuladas com os tópicos gerados pelo modelo através do algoritmo Gibbs Sampling. Erros de ortografia



(a)



(b)



(c)

Figura 18: Acurácia média do modelo Nota-Aspecto utilizando diferentes classificadores. (a) Máxima Entropia. (b) SVM. (c) Modelo Nota-Aspecto

foram mantidos propositadamente. Os índices F, S, A, V e O indicam que a palavra foi rotulada como aspecto *comida*, *serviço*, *ambiente*, *valor* ou *experiência*, respectivamente, e os sinais + e − indicam avaliações positivas e negativas, respectivamente, com relação ao aspecto escolhido. O índice NRA denota que a palavra foi assinalada a qualquer outro aspecto não subjetivo. Palavras não rotuladas foram removidas durante a etapa de pré-processamento.

Comida:+, Serviço:+, Ambiente:-, Valor:-, Experiência:-

the food^{O-} was actually^{NRA} incredible^{F-} in taste^{NRA}. Tha ratio of portion^{NRA} size^{NRA} to cost^{S-} was out of balance^{NRA}. I thought^{NRA} the portion^{V+} size^{NRA} could have been bigger^{NRA}. I left^{V-} hungry Our waitress^{S-} was a delight^{F-}, she was great^{F+}. Id consider^{NRA} going^{A-} back if there was a change^{V+} in portion^{NRA} size^{NRA}.

Comida:-, Serviço:-, Ambiente:-, Valor:-, Experiência:-

The food^{NRA} at the buffet^{S+} was terrible^{F-}. The quality^{NRA} of food^{O-} is better^{NRA} at Golden^{NRA} Corral and its over \$10^{NRA} cheaper^{S-}. There were only 4^{NRA} other parties^{NRA} in the restaurant^{NRA} when we were there which was on a thursday^{A+} night^{A-} at 6:30pm.

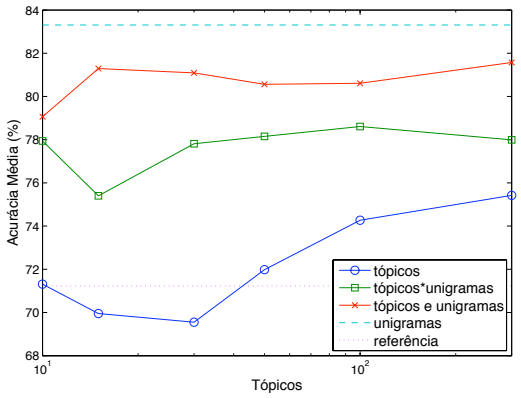
Figura 19: Resenhas cujas palavras foram rotuladas pelo modelo Nota-Aspecto (30 tópicos) utilizando as respectivas avaliações. Os índices F, S, A, V e O indicam que a palavra foi rotulada como aspecto comida, serviço, ambiente, valor ou experiência, respectivamente, e os sinais + e − indicam avaliações positivas e negativas, respectivamente, com relação ao aspecto escolhido. O índice NRA denota que a palavra foi assinalada a qualquer outro aspecto não subjetivo. Palavras não rotuladas foram removidas durante a etapa de pré-processamento.

Como é possível notar das amostra de resenhas rotuladas, as estimativas de rótulos apresentam resultando bastante insatisfatório, mesmo se desconsiderando a ordem e a palavra rotulada.

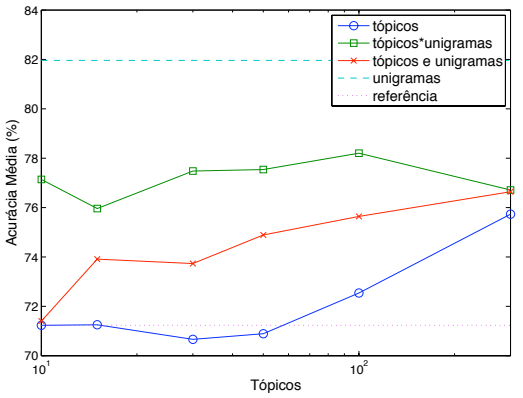
Modelo Nota-Aspecto LDA O modelo Nota-Aspecto LDA estende o modelo Nota-Aspecto permitindo distribuições *a priori* distintas para cada documento. Os resultados da utilização dos rótulos estimados na classificação de sentimentos para os classificadores MaxEnt, SVM e Nota-Aspecto LDA são exibidos na Figura 20.

Os resultados indicam que a inclusão das distribuições de mistura de tópicos não resulta em melhora da estimação dos rótulos em relação à discriminação de sentimentos. O desempenho, assim como no modelo Nota-Aspecto, é bastante insatisfatório.

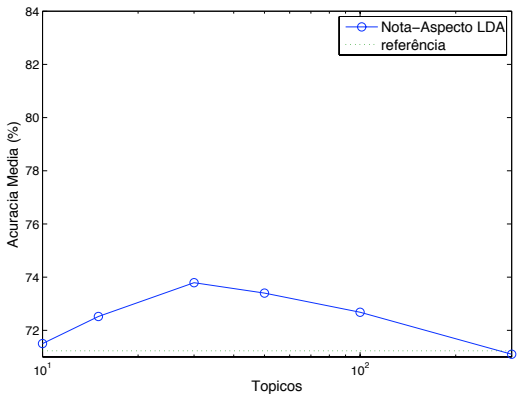
A Figura 21 exibe algumas amostras de resenhas rotuladas com os tópicos gerados pelo modelo através do algoritmo Gibbs Sampling. Mais uma vez é possível notar que, apesar da estimação de tópicos ser aparentemente satisfatória, a estimação de rótulos apresenta



(a)



(b)



(c)

Figura 20: Acurácia média do modelo Nota-Aspecto LDA utilizando diferentes classificadores. (a) Máxima Entropia. (b) SVM. (c) Modelo Nota-Aspecto LDA

um desempenho ruim, confirmando os resultados da análise qualitativa das estimativas de representação de documentos.

Comida:+, Serviço:+, Ambiente:-, Valor:-, Experiência:-

the food^{NRA} was actually^{V+} incredible^{F-} in taste^{NRA}. Tha ratio of portion^{NRA} size^{V+} to cost^{S-} was out of balance^{NRA}. I thought^{NRA} the portion^{NRA} size^{V+} could have been bigger^{V+}. I left^{V-} hungry Our waitress^{V-} was a delight^{F-}, she was great^{NRA}. Id consider^{NRA} going^{O+} back if there was a change^{O-} in portion^{NRA} size^{V+}.

Comida:-, Serviço:-, Ambiente:-, Valor:-, Experiência:-

The food^{NRA} at the buffet^{NRA} was terrible^{F-}. The quality^{NRA} of food^{NRA} is better^{F-} at Golden^{NRA} Corral and its over \$10^{S-} cheaper^{NRA}. There were only 4^{NRA} other parties^{NRA} in the restaurant^{A+} when we were there which was on a thursday^{NRA} night^{NRA} at 6:30pm.

Figura 21: Resenhas cujas palavras foram rotuladas pelo modelo Nota-Aspecto LDA (30 tópicos) utilizando as respectivas avaliações. Erros de ortografia foram mantidos propositadamente. Os índices F, S, A, V e O indicam que a palavra foi rotulada como aspecto comida, serviço, ambiente, valor ou experiência, respectivamente, e os sinais $+$ e $-$ indicam avaliações positivas e negativas, respectivamente, com relação ao aspecto escolhido. O índice NRA denota que a palavra foi assinalada a qualquer outro aspecto não subjetivo. Palavras não rotuladas foram removidas durante a etapa de pré-processamento.

5.5 Discussão

Os resultados obtidos demonstram que os modelos de tópicos analisados neste trabalho não contribuem significativamente para o aumento de acurácia da tarefa de classificação de sentimentos de resenhas de usuário no conjunto de dados testado. A maior parte dos modelos e configurações de atributos analisados apresentou desempenho tão bom ou pior que o da representação de unigramas tradicionalmente adotada na literatura. De fato, apenas três configurações obtiveram resultados acima do estado-da-arte (modelo LDA e de temas a representação de tópicos e unigramas e modelo LDA na representação tópico-unigramas). Os resultados apontam, acima de tudo, a importância de estimativas mais confiáveis para as distribuições *a priori* dos modelos.

As Tabelas 6, 7 e 8 apresentam os melhores resultados obtidos com as representações de tópicos, tópicos*unigramas e tópicos e unigramas, respectivamente, para o classificador MaxEnt. Os resultados obtidos com o classificador SVM ficaram abaixo do classificador MaxEnt e foram omitidos por não apresentar informação relevante. As tabelas apresentam os resultados de acurácia em relação a cada aspecto individualmente e a acurácia média (todos em porcentagem). Os resultados para o modelo LDA e para o modelo de temas foram obtidos com valores de α e β estimados pelo método de ponto fixo. Para os demais

modelos, quando necessário, foram utilizados os valores adotados na literatura ($\alpha = 50/T$, $\beta = 0,1$).

Tabela 6: Melhores resultados obtidos com o classificador de Máxima Entropia para cada modelo testado utilizando apenas os rótulos de tópicos na representação final. Os valores indicam acurácias em porcentagem.

Modelo	***T***	**Comida**	**Serviço**	**Valor**	**Ambiente**	**Experiência**	**Média**
LDA	50	86,50	81,15	79,95	82,03	85,94	**83,12**
Digrama	15	75,16	67,97	68,77	68,21	69,73	**69,97**
Temas	50	85,30	81,79	80,89	82,11	85,78	**83,18**
N-A	10	74,44	65,49	69,73	65,41	67,33	**68,48**
N-A LDA	300	78,75	75,08	71,41	74,04	77,79	**75,42**

Tabela 7: Melhores resultados obtidos com o classificador de Máxima Entropia para cada modelo testado utilizando a multiplicação entre rótulos de tópicos e unigramas na representação final. Os valores indicam acurácias em porcentagem.

Modelo	***T***	**Comida**	**Serviço**	**Valor**	**Ambiente**	**Experiência**	**Média**
LDA	10	86,42	80,51	79,31	82,98	87,86	**83,42**
Digrama	10	82,67	77,72	76,12	77,87	82,67	**79,41**
Temas	10	84,50	80,91	79,23	82,35	85,62	**82,52**
N-A	10	82,75	76,76	73,80	77,16	82,35	**78,56**
N-A LDA	100	81,31	76,52	74,12	77,56	80,43	**78,61**

A Tabela 9 exibe os resultados para os classificadores MaxEnt e SVM utilizando a representação de unigramas (*bag of words*) tradicional. A comparação desses resultados com os demonstrados nas Tabelas 6, 7 e 8 evidencia o pequeno ganho que as representações dos modelos de tópicos proporcionam na classificação de sentimentos nos dados utilizados.

Diversos fatores podem explicar a falta de sucesso da aplicação dos modelos de tópicos na construção de representações de resenhas de usuário com maior poder discriminatório dos sentimentos expressos. Um deles refere-se ao caráter não supervisionado dos algoritmos. De maneira geral, os métodos não supervisionados apresentam desempenho inferior ao dos métodos supervisionados (BLEI et al., 2003). Dessa forma, os valores estimados pelos modelos de tópicos contém incertezas e ruídos que quando utilizados indiscriminadamente podem levar à queda de desempenho do classificador. Vale notar que mesmo nos modelos que incorporam informação categórica, os métodos podem ser vistos como não supervisionados, pois as categorias fornecidas não são utilizadas de maneira supervisionada na estimação dos parâmetros. Considerando-se o caráter não supervisionado, as

Tabela 8: Melhores resultados obtidos com o classificador de Máxima Entropia para cada modelo testado utilizando rótulos de tópicos e unigramas na representação final. Os valores indicam acurácias em porcentagem.

Modelo	T	**Comida**	**Serviço**	**Valor**	**Ambiente**	**Experiência**	**Média**
LDA	15	86,66	82,67	79,71	83,71	86,74	**83,90**
Digrama	30	86,26	81,95	78,27	82,51	86,18	**83,04**
Temas	15	87,69	81,39	78,67	82,03	86,66	**83,29**
N-A	15	85,54	80,99	77,56	82,59	85,70	**82,48**
N-A LDA	300	85,62	78,35	75,56	81,79	86,50	**81,57**

Tabela 9: Resultados obtidos na classificação de sentimento utilizando a representação de unigramas (*bag of words*). Os valores indicam acurácias em porcentagem.

Classificador	**Comida**	**Serviço**	**Valor**	**Ambiente**	**Experiência**	**Média**
MaxEnt	85,86	81,55	78,83	82,83	87,46	**83,31**
SVM	84,82	80,91	77,72	80,83	85,30	**81,92**

representações de tópicos estimadas podem ser encarados como bem sucedidas, pois apresentam desempenho competitivo com as representações tradicionais que não se utilizam de uma etapa de estimação de dados não supervisionada.

Um outro fator possível para os resultados observados é o da estimação de tópicos do modelo LDA. Como notado por Titov e McDonald (2008b), os tópicos extraídos para conjunto de resenhas tendem a relacionar-se às diferentes classes de resenhas utilizadas, ao invés de capturar as diferentes opiniões existentes. Este resultado não é surpreendente, pois o modelo LDA foi originalmente desenvolvido para estimar os diferentes temas ou assuntos presentes em uma coleção heterogênea de textos, como na tradicional tarefa de categorização de texto. No entanto, na tarefa de classificação de sentimentos, o foco não é na discriminação dos assuntos tratados no texto (e.g., hotéis, restaurantes, bares) mas sim as possíveis informações subjetivas expressas. Uma inspeção manual realizada em algumas centenas de resenhas pelo autor deste trabalho constatou que existem 11 aspectos freqüentemente avaliados pelos autores no conjunto de dados utilizado. São eles: Qualidade da comida, preço da comida, variedade da comida, qualidade do serviço, agilidade do serviço, localização, limpeza, decoração, ruído, lotação e experiência. Nenhum desses aspectos aparece, de maneira clara, nos tópicos estimados pelos modelos LDA e de temas.

Os modelos Nota-Aspecto e Nota-Aspecto LDA resolvem parcialmente o problema da extração de tópicos representativos de sentimentos incorporando informação categórica

explicitamente na descrição dos tópicos, forçando-os, dessa forma, a descreverem as opiniões expressas em relação a aspectos particulares ao invés de temas gerais. No entanto, esses modelos não são capazes de beneficiarem-se desses tópicos para representar corretamente os documentos, levando a desempenhos muito aquém das técnicas tradicionais.

Por fim, é interessante ressaltar que os resultados obtidos também indicam que os modelos de tópicos, quando apropriadamente ajustados, são capazes de representar de forma sucinta e sem grande perda de informação as informações contidas no texto capturadas pelos modelos tradicionais de representação de texto. Dessa forma, os modelos de tópicos fornecem uma ferramenta valiosa para a análise qualitativa de coleções de texto em um espaço de dimensionalidade reduzida e fácil interpretação através dos vetores de tópicos e misturas de tópicos. As amostras de tópicos e de representações de documentos apresentam duas possíveis formas de extrair informação qualitativa de grandes coleções de texto.

6 CONCLUSÃO

A internet possibilita que um grande número de pessoas torne-se produtora e consumidora de informações antes escassas. O advento de novas tecnologias e aplicações tem incentivado de forma crescente usuários a participarem na elaboração de conteúdo, o que faz com que a quantidade de informação disponível ultrapasse a capacidade humana de gerenciamento.

A intensa pesquisa em recuperação de informação resultou no desenvolvimento de mecanismos de busca confiáveis e ágeis que "vasculham" a rede por informações e as organizam de forma a permitir seu acesso aos usuários (MANNING; RAGHAVAN; SCHÜTZE, 2008). No entanto, até o presente momento, grande parte dessa pesquisa tem focado no desenvolvimento de técnicas que auxiliem no gerenciamento de informações factuais, em grande parte, retornando ponteiros para fontes que contém textos de caráter informativo.

Informações factuais como artigos de enciclopédias, tutoriais e manuais *online*, artigos científicos e artigos de jornal constituem apenas parte do conteúdo disponível na internet. Em especial, uma grande parte do conteúdo gerado por usuários encontra-se na forma de opiniões, avaliações ou anotações pessoais. Por exemplo, em (MAUÁ; COZMAN, 2008b) e (MAUÁ; COZMAN, 2008a) os autores estudam aplicações para informações sobre relações de confiança disponibilizadas por usuários de forma a aumentar a confiabilidade do sistema e atuar como mecanismo de filtragem de dados. Sítios como *epinions* (`http://www.epinions.com`) encorajam usuários a avaliar o conteúdo postado por outros usuários, gerando redes de relacionamento que ajudam a garantir a confiabilidade dos dados postados e a superar a sobrecarga de informação. Sítios de leilão como *ebay* (`http://www.ebay.com`) e *Amazon* (`http://www.amazon.com`) utilizam as avaliações feitas pelos usuários em transações como mecanismo de reputação.

Uma parcela significativa das opiniões publicadas por usuários está na forma de resenhas (BLEI; LAFFERTY, 2009; PANG; LEE, 2008). Sítios especializados como C|net (`http://www.cnet.com`) e *Yelp* (`http://www.yelp.com`) permitem que usuários publi-

quem suas avaliações sobre produtos e serviços diversos. Um número crescente de sítios como lojas virtuais, *blogs* e portais governamentais e corporativos agora incentivam a interação com os usuários através de campos para a entrada de opiniões e críticas.

Este trabalho objetivou a aplicação de modelos de tópicos na tarefa de classificação de sentimentos baseada em múltiplos aspectos. O objetivo da classificação de sentimentos é superar a atual sobrecarga de informações criada pela abundância de conteúdo opinativo publicado por usuários na internet representando concisamente um documento pelos sentimentos expressos. De maneira geral, a classificação de sentimentos pode ser vista como uma subtarefa da categorização de texto, onde documentos são discriminados em relação ao assunto ou tópico a que pertençam. No entanto, as técnicas de classificação de sentimentos apresentam desempenho muito inferior aos classificadores de texto tradicionais.

A idéia por trás da utilização de modelos de tópicos na representação de resenhas de usuário é a de que documentos possam ser representados por distribuições de tópicos que se relacionam com um sentimento expresso no texto. Dessa forma, é possível superar problemas encontrados na representação tradicional de textos como ambigüidade léxica e dimensionalidade elevada.

No entanto, reforçando as observações de Titov e McDonald (2008b), os experimentos conduzidos com resenhas extraídas da internet, indicaram que o modelo LDA tradicional falha em extrair tópicos representativos das avaliações dos múltiplos aspectos presente nos textos. No seu lugar, o modelo LDA infere tópicos relacionados a categorias macro tais como o tipo de restaurante (e.g., comida indiana, comida italiana) e o tipo de informação (e.g., avaliação, descrição). A inclusão de informação de contexto de palavra e sentimento também não resultaram e melhoras significativas frente ao objetivo.

Dois modelos foram então desenvolvidos pelo autor de forma a garantir que os tópicos estimados sejam representativos das opiniões presentes. Os resultados apresentados demonstram que os modelos Nota-Aspecto e Nota-Aspecto LDA são capazes de extrair tópicos discriminativos dos sentimentos avaliados na coleção.

Apesar de bem sucedidos na tarefa de estimação de tópicos, os modelos Nota-Aspecto e Nota-Aspecto LDA falham em prover uma representação capaz de discriminar entre os sentimentos expressos no texto, levando a um mau desempenho na tarefa de classificação de sentimentos.

De maneira geral, os experimentos aqui relatados indicam que o uso de modelos de tópico na modelagem de resenhas de usuário não aumenta significativamente o desempenho

dos classificadores na tarefa de classificação de sentimento. Por outro lado, os modelos de tópicos demonstraram ser ferramentas úteis na análise de grandes quantidades de dados textuais, através da extração de tópicos e da representação de baixa dimensionalidade dos documentos. Por fim, cabe notar que por serem modelos probabilísticos generativos, os modelos estimados são facilmente interpretáveis.

REFERÊNCIAS

ANDRIEU, C. et al. An introduction to MCMC for machine learning. *Machine Learning*, v. 50, n. 1-2, p. 5–43, 2003.

ASUNCION, A. et al. On smoothing and inference for topic models. In: BILMES, J.; NG, A. (Ed.). *UAI '09: Proceedings of the 25th Conference in Uncertainty in Artificial Intelligence*. Quebec, Canada: AUAI Press, 2009.

BHATTACHARYA, I.; GETOOR, L. A latent Dirichlet model for unsupervised entity resolution. In: *SIAM International Conference on Data Mining*. Bethesda, MD, USA: [s.n.], 2006. p. 47–58. Disponível em: <http://www.cs.umd.edu/∼indrajit/DOCS/sdm06.pdf>.

BLEI, D.; LAFFERTY, J. Topic models. In: SRIVASTAVA, A.; SAHAMI, M. (Ed.). *Text Mining: Theory and Applications*. [S.l.]: Taylor and Francis, 2009.

BLEI, D.; MCAULIFFE, J. Supervised topic models. In: PLATT, J. C. et al. (Ed.). *Advances in Neural Information Processing Systems 20*. Cambridge, MA: MIT Press, 2008. p. 121–128.

BLEI, D. M.; JORDAN, M. I. Modeling annotated data. In: *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. New York, NY, USA: ACM, 2003. p. 127–134. ISBN 1-58113-646-3.

BLEI, D. M. et al. Latent Dirichlet allocation. *Journal of Machine Learning Research*, v. 3, p. 993–1022, 2003.

DAVE, K.; LAWRENCE, S.; PENNOCK, D. M. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: *WWW '03: Proceedings of the 12th International Conference on World Wide Web*. New York, NY, USA: ACM, 2003. p. 519–528. ISBN 1-58113-680-3.

DING, X.; LIU, B.; YU, P. S. A holistic lexicon-based approach to opinion mining. In: *WSDM '08: Proceedings of the International Conference on Web Search and Web Data Mining*. New York, NY, USA: ACM, 2008. p. 231–240. ISBN 978-1-59593-927-9.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple Bayesian classifier under zero-one loss. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 29, n. 2-3, p. 103–130, 1997. ISSN 0885-6125.

GILKS, W. R. *Markov Chain Monte Carlo in Practice*. [S.l.]: Chapman & Hall/CRC, 1995. ISBN 0412055511.

GRIFFITHS, T. L.; STEYVERS, M. Finding scientific topics. *Proc. Natl. Acad. Sci. USA*, v. 101 Suppl 1, p. 5228–5235, April 2004. ISSN 0027-8424.

GRIFFITHS, T. L. et al. Integrating topics and syntax. In: *In Advances in Neural Information Processing Systems 17*. [S.l.]: MIT Press, 2005. p. 537–544.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H. *The Elements of Statistical Learning*. 3. ed. Canada: Springer, 2001. ISBN 0387952845.

HOFMANN, T. Probabilistic latent semantic indexing. In: *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 1999. p. 50–57. ISBN 1-58113-096-1.

JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: *ECML '98: Proceedings of the 10th European Conference on Machine Learning*. London, UK: Springer-Verlag, 1998. p. 137–142.

JOACHIMS, T. Making large-scale support vector machine learning practical. *Advances in kernel methods: support vector learning*, MIT Press, Cambridge, MA, USA, p. 169–184, 1999.

LACOSTE-JULIEN, S.; SHA, F.; JORDAN, M. DiscLDA: Discriminative learning for dimensionality reduction and classification. In: KOLLER, D. et al. (Ed.). *NIPS '08: Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems*. Vancouver, Canada: MIT Press, 2008. p. 897–904. Disponível em: <http://books.nips.cc/papers/files/nips21/NIPS2008_0993.pdf>.

LI, F.-F.; PERONA, P. A Bayesian hierarchical model for learning natural scene categories. In: *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Washington, DC, USA: IEEE Computer Society, 2005. v. 2, p. 524–531. ISBN 0-7695-2372-2.

MAATEN, L. van der. Visualizing data using t-SNE. *Journal of Machine Learning Research*, v. 9, p. 2579–2605, 2008.

MALOUF, R. A comparison of algorithms for maximum entropy parameter estimation. In: *COLING '02: Proceedings of the 6th Conference on Natural Language Learning*. Morristown, NJ, USA: Association for Computational Linguistics, 2002. p. 1–7.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. *Introduction to Information Retrieval*. Cambridge, MA, USA: Cambridge University Press, 2008. ISBN 0521865719.

MANNING, C. D.; SCHÜTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

MAUÁ, D.; COZMAN, F. Managing trust in virtual communities with markov logic. In: *WTDIA '08: IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence*. Salvador, Brazil: [s.n.], 2008.

MAUÁ, D.; COZMAN, F. Using social data to predict trust on web communities: A case study with the epinions.com website. In: *WIVA '08: Anais do Workshop on Information Visualization and Analysis in Social Networks*. Campinas, Brazil: [s.n.], 2008. p. 1–10.

MAUÁ, D.; COZMAN, F. Representing and classifying user reviews. In: *ENIA '09: VIII Encontro Nacional de Inteligência Artificial*. Brazil: [s.n.], 2009.

MIMNO, D. M.; MCCALLUM, A. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In: MCALLESTER, D. A.; MYLLYMÄKI, P. (Ed.). *UAI '08: Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*. Helsinki, Finland: AUAI Press, 2008. p. 411–418. ISBN 0-9749039-4-9.

MINKA, T. P. Estimating a Dirichlet distribution. 2003. Disponível em: <http://research.microsoft.com/∼minka>.

MINKA, T. P.; LAFFERTY, J. D. Expectation-propogation for the generative aspect model. In: DARWICHE, A.; FRIEDMAN, N. (Ed.). *UAI '02, Proceedings of the 18th Conference in Uncertainty in Artificial Intelligence*. Alberta, Canada: Morgan Kaufmann, 2002. p. 352–359. ISBN 1-55860-897-4.

NIGAM, K. Using maximum entropy for text classification. In: *IJCAI '99: Workshop on Machine Learning for Information Filtering*. Stockholm, Sweden: [s.n.], 1999. p. 61–67.

PANG, B. Seeing stars. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL 05 ACL 05*. [S.l.: s.n.], 2005. p. 115.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1-2, p. 1–135, 2008. ISSN 1554-0669.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: *EMNLP '02: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2002. p. 79–86.

PORTEOUS, I. et al. Fast collapsed gibbs sampling for latent Dirichlet allocation. In: *KDD '08: Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2008. p. 569–577. ISBN 978-1-60558-193-4.

RICHARDSON, M.; DOMINGOS, P. Markov logic networks. *Machine Learning*, Springer, v. 62, n. 1-2, p. 107–136, February 2006. ISSN 0885-6125.

ROSEN-ZVI, M. et al. The author-topic model for authors and documents. In: CHICKERING, D. M.; HALPERN, J. Y. (Ed.). *AUAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*. Arlington, Virginia, USA: AUAI Press, 2004. p. 487–494. ISBN 0-9749039-0-6.

SEBAH, P.; GOURDON, X. Introduction to the Gamma function. 2002.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Comput. Surv.*, ACM Press, v. 34, n. 1, p. 1–47, March 2002. ISSN 0360-0300.

SIVIC, J. et al. Discovering object categories in image collections. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. [S.l.: s.n.], 2005.

SNYDER, B.; BARZILAY, R. Multiple aspect ranking using the good grief algorithm. In: *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Rochester, New York: Association for Computational Linguistics, 2007. p. 300–307. Disponível em: <http://www.aclweb.org/anthology/N/N07/N07-1038>.

STEYVERS, M.; GRIFFITHS, T. Probabilistic topic models. In: LANDAUER, T. et al. (Ed.). *Handbook of Latent Semantic Analysis*. London, UK: Lawrence Erlbaum Associates, 2007. ISBN 1410615340.

TITOV, I.; MCDONALD, R. A joint model of text and aspect ratings for sentiment summarization. In: *Human Language Technologies 2008: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. Columbus, Ohio: Association for Computational Linguistics, 2008. p. 308–316. Disponível em: <http://www.aclweb.org/anthology/P/P08/P08-1036>.

TITOV, I.; MCDONALD, R. Modeling online reviews with multi-grain topic models. In: *WWW '08: Proceedings of the 17th International Conference on World Wide Web*. New York, NY, USA: ACM, 2008. p. 111–120. ISBN 978-1-60558-085-2.

TURNEY, P. D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Morristown, NJ, USA: Association for Computational Linguistics, 2001. p. 417–424.

WALLACH, H. M. Topic modeling: beyond bag-of-words. In: *ICML '06: Proceedings of the 23rd international conference on Machine learning*. New York, NY, USA: ACM, 2006. p. 977–984. ISBN 1-59593-383-2.

XING, D. Employing latent Dirichlet allocation for fraud detection in telecommunications. *Pattern Recognition Letters*, v. 28, n. 13, p. 1727, 2007.

YU, H.; HATZIVASSILOGLOU, V. Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: *EMNL '03: Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Morristown, NJ, USA: Association for Computational Linguistics, 2003. p. 129–136.

APÊNDICE A – DERIVAÇÃO DO ALGORITMO GIBBS SAMPLING PARA O MODELO LDA

O algoritmo Gibbs Sampling aproxima a distribuição *a posteriori* tomando amostras das distribuições marginais condicionais nas demais variáveis latentes dadas por

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) \propto P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i})P(z_i = j|\mathbf{z}_{-i}) \; . \tag{A.1}$$

O primeiro termo no lado direito da Equação (A.1) pode ser rescrito como

$$P(w_i|z_i = j, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \int_{\varphi_j} P(w_i|z_i = j, \varphi_j)P(\varphi_j|\mathbf{z}_{-i}, \mathbf{w}_{-i})\, d\varphi_j \; , \tag{A.2}$$

onde $P(w_i = v_k|z_i = j, \varphi_j) = \varphi_{j_k}$ é a probabilidade da palavra $w_i = v_k$ de acordo com o tópico φ_j.

Denotando por $n^{(k)}_{-i,j'} = \sum_{i'=1,i'\neq i}^{N} \mathbf{1}(w_{i'} = v_k, z_{i'} = j')$ o número de palavras assinaladas ao j'-ésimo tópico não incluindo a i-ésima palavra, pode-se rescrever o termo mais à direita na Equação (A.2) como

$$\begin{aligned}
P(\varphi_j|\mathbf{z}_{-i}, \mathbf{w}_{-i}) &= \int_{\varphi_{-j}} P(\varphi|\mathbf{z}_{-i}, \mathbf{w}_{-i})\, d\varphi_{-j} \\
&= \tfrac{1}{P(\mathbf{w}_{-i}|\mathbf{z}_{-i})} \int_{\varphi_{-j}} P(\mathbf{w}_{-i}|\mathbf{z}_{-i}, \varphi)P(\varphi)\, d\varphi_{-j} \\
&= \tfrac{1}{P(\mathbf{w}_{-i}|\mathbf{z}_{-i})} \int_{\varphi_{-j}} \left(\textstyle\prod_{i'=1,i'\neq i}^{N} P(w_{i'}|z_{i'}, \varphi)\right) \left(\textstyle\prod_{j'=1}^{T} P(\varphi_{j'})\right)\, d\varphi_{-j} \\
&= \tfrac{B}{P(\mathbf{w}_{-i}|\mathbf{z}_{-i})} \int_{\varphi_{-j}} \textstyle\prod_{j'=1}^{T} \prod_{k=1}^{W} \varphi_{j'_k}^{n^{(k)}_{-i,j'}+\beta_k-1}\, d\varphi_{-j} \\
&= \left(\textstyle\prod_{k=1}^{W} \varphi_{j_k}^{n^{(k)}_{-i,j}+\beta_k-1}\right) \tfrac{B}{P(\mathbf{w}_{-i}|\mathbf{z}_{-i})} \left(\textstyle\prod_{j'=1,j'\neq j}^{T} \int_{\varphi_{j'}} \prod_{k=1}^{W} \varphi_{j'_k}^{n^{(k)}_{-i,j'}+\beta_k-1}\, d\varphi_{j'}\right) \; .
\end{aligned} \tag{A.3}$$

O denominador $P(\mathbf{w}_{-i}|\mathbf{z}_{-i})$ pode ser deduzido da mesma forma que a Equação (3.3),

$$P(\mathbf{w}_{-i}|\mathbf{z}_{-i}) = B \prod_{j=1}^{T} \frac{\prod_{k=1}^{W} \Gamma(n^{(k)}_{-i,j} + \beta_k)}{\Gamma(\sum_k n^{(k)}_{-i,j} + \beta_k)} \; .$$

Cancelando termos comuns no numerador e denominador chega-se a

$$P(\varphi_j|\mathbf{z}_{-i},\mathbf{w}_{-i}) = \frac{\prod_{k=1}^{W}\varphi_{j_k}^{n_{-i,j}^{(k)}+\beta_k-1}}{\int_{\varphi_j}\prod_{k=1}^{W}\varphi_{j_k}^{n_{-i,j}^{(k)}+\beta_k-1}\,d\varphi_j} = \text{Dirichlet}(n_{-i,j}+\beta)\ . \tag{A.4}$$

Assim, a Equação (A.2) torna-se simplesmente a esperança marginal de uma distribuição Dirichlet dada por (MINKA, 2003)

$$P(w_i = v_k|z_i = j,\mathbf{z}_{-i},\mathbf{w}_{-i}) = \frac{n_{-i,j}^{(k)}+\beta_k}{\sum_k n_{-i,j}^{(k)}+\beta_k}\ . \tag{A.5}$$

O termo mais à direita na Equação (A.1) pode ser obtido de forma semelhante. Primeiro, denote por d o índice do documento ao qual pertence a i-ésima palavra w_i. Então, integrando em θ_d chega-se a

$$P(z_i = j|\mathbf{z}_{-i}) = \int_{\theta_d} P(z_i = j|\theta_d)P(\theta_d|\mathbf{z}_{-i})\,d\theta_d\ . \tag{A.6}$$

Por definição, $P(z_i = j|\theta_d) = \theta_{d_j}$ é a probabilidade do rótulo de tópico de acordo com a distribuição de tópicos θ_d do d-ésimo documento. O segundo termo pode ser rescrito como

$$P(\theta_d|\mathbf{z}_{-i}) = \frac{P(\mathbf{z}_{-\mathbf{i}}|\theta_d)P(\theta_d)}{P(\mathbf{z}_{-\mathbf{i}})} = \frac{1}{P(\mathbf{z}_{-\mathbf{i}})}\int_{\theta_{-d}} P(\mathbf{z}_{-\mathbf{i}}|\theta)P(\theta)\,d\theta\ .$$

O denominador pode ser resolvido da mesma forma que a Equação (3.4). Cancelando termos comuns chega-se a

$$P(\theta_d|\mathbf{z}_{-i}) = \frac{\prod_{j=1}^{T}\theta_{d_j}^{n_{-i,j}^{(d)}+\alpha_j-1}}{\int_{\theta_d}\prod_{j=1}^{T}\theta_{d_j}^{n_{-i,j}^{(d)}+\alpha_k-1}\,d\theta_d} = \text{Dirichlet}(n_{-i,d}+\alpha)\ .$$

Novamente, a Equação (A.6) resume-se à esperança marginal de uma distribuição Dirichlet. Portanto,

$$P(z_i = j|\mathbf{z}_{-i}) = \frac{n_{-i,j}^{(d)}+\alpha_j}{\sum_j n_{-i,j}^{(d)}+\alpha_j}\ . \tag{A.7}$$

De (A.5) e (A.7), pode-se obter a probabilidade marginal a menos de um fator de escala

$$P(z_i = j|\mathbf{z}_{-i},\mathbf{w}) \propto \frac{n_{-i,j}^{(k)}+\beta_k}{\sum_k n_{-i,j}^{(k)}+\beta_k}\frac{n_{-i,j}^{(d)}+\alpha_j}{\sum_j n_{-i,j}^{(d)}+\alpha_j}\ . \tag{A.8}$$

A distribuição marginal $P(z_i|\mathbf{z}_{-i},\mathbf{w})$ pode ser obtida através da normalização da Equa-

ção A.8

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{w}) = \frac{1}{Z}\frac{(n^{(k)}_{-i,j} + \beta_k)(n^{(d)}_{-i,j} + \alpha_j)}{\sum_k n^{(k)}_{-i,j} + \beta_k}, \tag{A.9}$$

onde $Z = \sum_{j'=1}^{T} \frac{(n^{(k)}_{-i,j}+\beta_k)(n^{(d)}_{-i,j}+\alpha_j)}{\sum_k n^{(k)}_{-i,j}+\beta_k}$.

APÊNDICE B – DERIVAÇÃO DO ALGORITMO DE ESTIMAÇÃO DOS HIPERPARÂMETROS DO MODELO LDA

A idéia por trás do esquema de iteração de ponto fixo é semelhante à idéia por trás do algoritmo *Expectation-Maximization*. De forma a maximizar uma determinada função objetivo $F(x)$, o método consiste em encontrar um limite inferior $L(x) \leqslant F(x), \forall x$ e maximizar $L(x)$. Assim como no algoritmo EM, esse procedimento, de maneira geral, não garante convergência para o máximo global. No entanto, devido à distribuição multinomial-Dirichlet ser convexa, este procedimento aplicado ao modelo LDA garante que um máximo global seja encontrado (MINKA, 2003).

Um limite inferior à função de log verosimilhança pode ser obtido a partir das desigualdades da função gama encontradas na literatura

$$\log\Gamma(x) + \log\Gamma(n+x) \geq \log\Gamma(\hat{x}) + (\hat{x}-x)(\Psi(n+\hat{x}) - \Psi(\hat{x})) - \log\Gamma(n+\hat{x}), \quad \text{(B.1)}$$

e

$$\log\Gamma(n+x) - \log\Gamma(x) \geq \log\Gamma(n+\hat{x}) - \log\Gamma(\hat{x}) + (\Psi(n+\hat{x}) - \Psi(\hat{x}))\,\hat{x}\left(\log\left(\frac{x}{\hat{x}}\right)\right), \quad \text{(B.2)}$$

onde $\hat{x}$ é um ponto próximo a x, n é um número inteiro e Ψ é a função digama dada pela derivada da função log da função Gama, $\Psi = \log\Gamma'$.

Como exemplo, o termo $\log P(\mathbf{z}|\alpha)$ na Equação (3.11) pode ser rescrito como uma soma de log probabilidades, $\log P(\mathbf{z}|\alpha) = \sum_{d=1}^{D} \log P(\mathbf{z}_d|\alpha)$. Cada termo $\log P(\mathbf{z}_d|\alpha)$ pode, por vez, ser rescrito como

$$\begin{aligned} \log P(\mathbf{z}_d|\alpha) \;=\; & \underbrace{\log\Gamma(\alpha_0) - \log\Gamma(\sum_j n_j^{(d)} + \alpha_j)}_{1} \\ & + \sum_{j=1}^{T} \underbrace{\log\Gamma(n_j^{(d)} + \alpha_j) - \log\Gamma(\alpha_j)}_{2}\,. \end{aligned} \quad \text{(B.3)}$$

Utilizando a Equação (B.1) com $\alpha_j^{(i)}$ como x e $\alpha_j^{(i-1)}$ como $\hat{x}$, pode-se chegar a um limite inferior para o termo 1 na Equação (B.3)

$$\begin{aligned}\log\Gamma(\alpha_0) - \log\Gamma(\textstyle\sum_j n_j^{(d)} + \alpha_j) \geq \log\Gamma(\alpha_0^{(i-1)}) \\ +(\alpha_0^{(i-1)} - \alpha_0^{(i)})\left(\Psi(\textstyle\sum_j n_j^{(d)} + \alpha_j^{(i-1)}) - \Psi(\alpha_0^{(i-1)})\right) - \log\Gamma(\textstyle\sum_j n_j^{(d)} + \alpha_j^{(i-1)}) \ .\end{aligned} \tag{B.4}$$

De maneira similar, pode-se utilizar a Equação (B.2) para chegar a um limite inferior para o termo 2

$$\begin{aligned}&\log\Gamma(n_j^{(d)} + \alpha_j) - \log\Gamma(\alpha_j) \geq \log\Gamma(n_j^{(d)} + \alpha_j^{(i-1)}) - \log\Gamma(\alpha_j^{(i-1)}) \\ &+\left(\Psi(n_j^{(d)} + \alpha_j^{(i-1)}) - \Psi(\alpha_j^{(i-1)})\right)\alpha_j^{(i-1)}\left(\log\left(\frac{\alpha_j^{(i)}}{\alpha_j^{(i-1)}}\right)\right) \ .\end{aligned} \tag{B.5}$$

Um limite inferior para $\log P(\mathbf{z}|\alpha)$ pode ser encontrado por

$$L(\alpha) = \frac{1}{S}\sum_{s=1}^{S}\sum_{d=1}^{D} Eq.\ (B.4) + \sum_{j=1}^{T} Eq.\ (B.5) \ .$$

A equação de atualização para α pode ser obtida através da derivada do limite inferior acima. Um limite inferior para a componente $\log P(\mathbf{w}|\mathbf{z}, \beta)$ poder ser obtido de forma bastante similar.

ANEXO A – TÓPICOS ESTIMADOS

Neste anexo são apresentadas as listas completas de tópicos para os modelos LDA, Nota-Aspecto e Nota-Aspecto LDA "treinados" com 10, 15, 30 e 50 tópicos, $\alpha = 50/T$ e $\beta = 0,1$. Para o modelo de temas são apresentadas apenas as listas de modelos estimados com 10, 15 e 30 tópicos e com informações sobre o aspecto *comida* e *serviço*, omitindo-se as listas referentes aos aspectos *valor*, *ambiente* e *experiência*. A numeração dos tópicos apresentada nas tabelas é meramente ilustrativa, não havendo necessariamente relacionamento entre tópicos de mesmo índice em listas distintas.

Tabela 10: Tópicos estimados com o modelo **LDA** e **10 tópicos** no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	chicken$^{0.17}$ sauce$^{0.13}$ served$^{0.13}$ salad$^{0.13}$ shrimp$^{0.10}$ soup$^{0.09}$ fried$^{0.08}$ crab$^{0.07}$ cream$^{0.07}$ rice$^{0.07}$
1	food$^{1.04}$ service$^{0.37}$ good$^{0.36}$ restaurant$^{0.31}$ excellent$^{0.15}$ price$^{0.14}$ return$^{0.12}$ quality$^{0.11}$ visit$^{0.10}$ experience$^{0.09}$
2	did$^{0.16}$ table$^{0.14}$ minutes$^{0.13}$ waitress$^{0.12}$ ordered$^{0.11}$ told$^{0.11}$ waiter$^{0.11}$ asked$^{0.10}$ order$^{0.09}$ said$^{0.09}$
3	bar$^{0.31}$ it's$^{0.15}$ room$^{0.13}$ area$^{0.13}$ dining$^{0.12}$ good$^{0.11}$ nice$^{0.11}$ place$^{0.10}$ little$^{0.09}$ old$^{0.07}$
4	really$^{0.18}$ got$^{0.16}$ little$^{0.13}$ fries$^{0.12}$ came$^{0.11}$ sandwich$^{0.11}$ didn't$^{0.10}$ cheese$^{0.09}$ burger$^{0.09}$ pretty$^{0.08}$
5	great$^{0.96}$ food$^{0.60}$ atmosphere$^{0.30}$ friendly$^{0.28}$ staff$^{0.21}$ wonderful$^{0.21}$ place$^{0.18}$ recommend$^{0.17}$ service$^{0.16}$ family$^{0.13}$
6	menu$^{0.21}$ lunch$^{0.14}$ seafood$^{0.13}$ buffet$^{0.13}$ dishes$^{0.12}$ breakfast$^{0.09}$ fresh$^{0.09}$ items$^{0.07}$ cafe$^{0.07}$ variety$^{0.07}$
7	restaurant$^{0.37}$ experience$^{0.22}$ wine$^{0.17}$ dinner$^{0.17}$ steak$^{0.10}$ chef$^{0.09}$ dining$^{0.09}$ special$^{0.09}$ wife$^{0.08}$ list$^{0.07}$
8	time$^{0.24}$ don't$^{0.17}$ like$^{0.16}$ eat$^{0.14}$ know$^{0.14}$ say$^{0.13}$ people$^{0.13}$ just$^{0.13}$ going$^{0.12}$ bad$^{0.12}$
9	best$^{0.29}$ pizza$^{0.27}$ place$^{0.16}$ i've$^{0.14}$ new$^{0.14}$ like$^{0.13}$ years$^{0.12}$ home$^{0.10}$ town$^{0.10}$ italian$^{0.09}$

Tabela 11: Tópicos estimados com o modelo **LDA** e **15 tópicos** no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	lunch$^{0.20}$ breakfast$^{0.15}$ menu$^{0.13}$ cafe$^{0.11}$ local$^{0.09}$ coffee$^{0.09}$ open$^{0.08}$ located$^{0.08}$ large$^{0.07}$ sunday$^{0.07}$
1	good$^{0.41}$ pizza$^{0.40}$ it's$^{0.25}$ really$^{0.22}$ like$^{0.16}$ little$^{0.15}$ place$^{0.15}$ order$^{0.12}$ pretty$^{0.10}$ places$^{0.10}$
2	ordered$^{0.27}$ food$^{0.17}$ husband$^{0.16}$ time$^{0.13}$ 2$^{0.13}$ steak$^{0.09}$ cold$^{0.09}$ better$^{0.09}$ water$^{0.08}$ 3$^{0.08}$
3	great$^{1.53}$ food$^{0.85}$ place$^{0.53}$ service$^{0.38}$ love$^{0.24}$ family$^{0.21}$ wonderful$^{0.17}$ friends$^{0.17}$ atmosphere$^{0.16}$ eat$^{0.14}$
4	best$^{0.71}$ new$^{0.27}$ try$^{0.23}$ visit$^{0.21}$ years$^{0.21}$ favorite$^{0.17}$ i've$^{0.14}$ times$^{0.12}$ home$^{0.10}$ area$^{0.10}$
5	table$^{0.22}$ minutes$^{0.21}$ did$^{0.20}$ waitress$^{0.18}$ told$^{0.15}$ said$^{0.15}$ waiter$^{0.14}$ asked$^{0.14}$ order$^{0.12}$ took$^{0.12}$
6	don't$^{0.28}$ time$^{0.23}$ just$^{0.23}$ eat$^{0.21}$ like$^{0.21}$ people$^{0.20}$ know$^{0.20}$ want$^{0.18}$ going$^{0.17}$ way$^{0.17}$
7	salad$^{0.18}$ served$^{0.17}$ sauce$^{0.12}$ cream$^{0.11}$ soup$^{0.10}$ dessert$^{0.09}$ chocolate$^{0.08}$ pasta$^{0.08}$ entrees$^{0.08}$ cheese$^{0.07}$
8	bar$^{0.49}$ good$^{0.44}$ night$^{0.20}$ area$^{0.18}$ nice$^{0.15}$ room$^{0.13}$ dinner$^{0.11}$ table$^{0.10}$ quite$^{0.10}$ time$^{0.09}$
9	food$^{0.78}$ excellent$^{0.46}$ service$^{0.43}$ friendly$^{0.41}$ staff$^{0.37}$ atmosphere$^{0.26}$ prices$^{0.23}$ recommend$^{0.19}$ good$^{0.16}$ wait$^{0.15}$
10	dishes$^{0.20}$ seafood$^{0.20}$ buffet$^{0.19}$ fresh$^{0.19}$ shrimp$^{0.19}$ sushi$^{0.16}$ fish$^{0.15}$ crab$^{0.14}$ menu$^{0.13}$ rice$^{0.11}$
11	restaurant$^{0.49}$ experience$^{0.24}$ food$^{0.22}$ service$^{0.21}$ owner$^{0.14}$ party$^{0.10}$ business$^{0.09}$ customers$^{0.09}$ wife$^{0.08}$ person$^{0.07}$
12	got$^{0.19}$ really$^{0.17}$ fries$^{0.16}$ came$^{0.15}$ sandwich$^{0.15}$ little$^{0.13}$ burger$^{0.12}$ didn't$^{0.10}$ wasn't$^{0.08}$ cheese$^{0.07}$
13	restaurant$^{0.34}$ wine$^{0.28}$ dining$^{0.26}$ experience$^{0.19}$ chef$^{0.14}$ menu$^{0.13}$ enjoyed$^{0.11}$ list$^{0.10}$ dinner$^{0.10}$ fine$^{0.10}$
14	chicken$^{0.42}$ like$^{0.17}$ hot$^{0.17}$ sauce$^{0.14}$ bbq$^{0.14}$ pork$^{0.12}$ beef$^{0.12}$ mexican$^{0.10}$ meat$^{0.09}$ just$^{0.08}$

Tabela 12: Tópicos estimados com o modelo **LDA** e **30 tópicos** no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	know$^{0.40}$ make$^{0.39}$ just$^{0.39}$ way$^{0.36}$ want$^{0.31}$ don't$^{0.30}$ sure$^{0.29}$ people$^{0.23}$ going$^{0.22}$ eat$^{0.10}$
1	breakfast$^{0.32}$ cafe$^{0.21}$ lunch$^{0.16}$ diner$^{0.15}$ coffee$^{0.14}$ day$^{0.12}$ local$^{0.12}$ sunday$^{0.10}$ hotel$^{0.09}$ eggs$^{0.09}$
2	restaurant$^{0.69}$ family$^{0.56}$ steak$^{0.43}$ special$^{0.28}$ visit$^{0.25}$ wife$^{0.22}$ tea$^{0.09}$ meal$^{0.09}$ feel$^{0.07}$ children$^{0.07}$
3	served$^{0.23}$ sauce$^{0.18}$ garlic$^{0.12}$ fresh$^{0.11}$ grilled$^{0.10}$ cheese$^{0.08}$ salad$^{0.07}$ mushrooms$^{0.06}$ entrees$^{0.06}$ potatoes$^{0.06}$
4	best$^{1.50}$ i've$^{0.37}$ try$^{0.27}$ favorite$^{0.25}$ town$^{0.17}$ times$^{0.14}$ area$^{0.12}$ eaten$^{0.11}$ we've$^{0.10}$ far$^{0.10}$
5	minutes$^{0.41}$ wait$^{0.31}$ table$^{0.27}$ seated$^{0.24}$ arrived$^{0.18}$ hour$^{0.17}$ order$^{0.16}$ waited$^{0.14}$ took$^{0.13}$ waiter$^{0.12}$
6	great$^{2.92}$ food$^{0.63}$ place$^{0.58}$ service$^{0.39}$ love$^{0.33}$ atmosphere$^{0.21}$ awesome$^{0.14}$ wonderful$^{0.12}$ fantastic$^{0.08}$ fabulous$^{0.06}$
7	fries$^{0.38}$ sandwich$^{0.37}$ burger$^{0.30}$ cheese$^{0.20}$ menu$^{0.11}$ burgers$^{0.10}$ lunch$^{0.10}$ bacon$^{0.09}$ sandwiches$^{0.07}$ really$^{0.07}$
8	manager$^{0.22}$ did$^{0.22}$ owner$^{0.18}$ told$^{0.17}$ asked$^{0.12}$ rude$^{0.11}$ service$^{0.10}$ money$^{0.10}$ experience$^{0.10}$ said$^{0.10}$
9	new$^{0.67}$ years$^{0.39}$ home$^{0.20}$ location$^{0.16}$ stop$^{0.14}$ restaurant$^{0.14}$ old$^{0.13}$ wings$^{0.11}$ ago$^{0.10}$ trip$^{0.10}$
10	food$^{0.78}$ service$^{0.52}$ restaurant$^{0.30}$ 2$^{0.29}$ time$^{0.29}$ people$^{0.24}$ times$^{0.22}$ eat$^{0.21}$ 3$^{0.19}$ party$^{0.14}$
11	waitress$^{0.43}$ came$^{0.32}$ did$^{0.24}$ ordered$^{0.23}$ didn't$^{0.20}$ said$^{0.19}$ meal$^{0.18}$ asked$^{0.16}$ table$^{0.16}$ got$^{0.15}$
12	pizza$^{0.79}$ good$^{0.21}$ cheese$^{0.20}$ order$^{0.18}$ chicago$^{0.15}$ italian$^{0.14}$ chain$^{0.11}$ style$^{0.11}$ just$^{0.09}$ crust$^{0.09}$
13	really$^{0.28}$ got$^{0.23}$ didn't$^{0.15}$ little$^{0.15}$ just$^{0.13}$ wasn't$^{0.12}$ pretty$^{0.10}$ there's$^{0.10}$ lot$^{0.10}$ came$^{0.08}$
14	food$^{1.11}$ excellent$^{0.61}$ service$^{0.49}$ prices$^{0.47}$ price$^{0.33}$ quality$^{0.28}$ reasonable$^{0.22}$ portions$^{0.17}$ atmosphere$^{0.14}$ return$^{0.12}$
15	chicken$^{0.57}$ bbq$^{0.30}$ pork$^{0.23}$ sauce$^{0.18}$ beef$^{0.18}$ fried$^{0.16}$ ribs$^{0.15}$ meat$^{0.14}$ sweet$^{0.12}$ meal$^{0.10}$
16	sushi$^{0.36}$ fresh$^{0.35}$ chef$^{0.26}$ fish$^{0.25}$ try$^{0.16}$ dishes$^{0.15}$ san$^{0.12}$ rolls$^{0.12}$ menu$^{0.11}$ prepared$^{0.10}$
17	food$^{0.42}$ mexican$^{0.23}$ restaurant$^{0.18}$ chips$^{0.13}$ like$^{0.12}$ authentic$^{0.11}$ chicken$^{0.10}$ salsa$^{0.09}$ tasty$^{0.08}$ portions$^{0.07}$
18	bar$^{0.93}$ area$^{0.18}$ beer$^{0.15}$ music$^{0.13}$ outside$^{0.12}$ seating$^{0.12}$ drinks$^{0.10}$ sit$^{0.09}$ pub$^{0.08}$ tables$^{0.08}$
19	dining$^{0.77}$ experience$^{0.44}$ restaurant$^{0.44}$ room$^{0.29}$ fine$^{0.15}$ beautiful$^{0.15}$ area$^{0.10}$ view$^{0.08}$ outstanding$^{0.06}$ setting$^{0.06}$
20	dessert$^{0.36}$ delicious$^{0.26}$ cream$^{0.26}$ chocolate$^{0.23}$ ice$^{0.18}$ perfect$^{0.14}$ cake$^{0.12}$ dinner$^{0.11}$ course$^{0.10}$ homemade$^{0.09}$
21	friendly$^{0.77}$ staff$^{0.67}$ wonderful$^{0.37}$ recommend$^{0.35}$ atmosphere$^{0.27}$ highly$^{0.26}$ excellent$^{0.23}$ food$^{0.21}$ experience$^{0.20}$ attentive$^{0.16}$
22	salad$^{0.60}$ ordered$^{0.46}$ bread$^{0.31}$ soup$^{0.22}$ plate$^{0.14}$ served$^{0.14}$ entree$^{0.09}$ cold$^{0.09}$ tasted$^{0.09}$ pasta$^{0.08}$
23	buffet$^{0.38}$ seafood$^{0.34}$ shrimp$^{0.32}$ crab$^{0.26}$ dishes$^{0.17}$ chinese$^{0.14}$ rice$^{0.12}$ lunch$^{0.11}$ variety$^{0.10}$ thai$^{0.09}$
24	good$^{2.34}$ nice$^{0.55}$ little$^{0.39}$ quite$^{0.24}$ really$^{0.21}$ food$^{0.21}$ pretty$^{0.18}$ bit$^{0.12}$ place$^{0.08}$ worth$^{0.07}$
25	dinner$^{0.43}$ time$^{0.36}$ husband$^{0.34}$ night$^{0.30}$ went$^{0.29}$ friends$^{0.17}$ friend$^{0.15}$ going$^{0.13}$ decided$^{0.13}$ tried$^{0.12}$
26	table$^{0.23}$ restaurant$^{0.22}$ server$^{0.19}$ water$^{0.14}$ kitchen$^{0.14}$ clean$^{0.10}$ servers$^{0.10}$ plates$^{0.08}$ establishment$^{0.07}$ patrons$^{0.07}$
27	menu$^{0.66}$ wine$^{0.56}$ list$^{0.21}$ italian$^{0.16}$ house$^{0.14}$ specials$^{0.13}$ dinner$^{0.13}$ large$^{0.10}$ offers$^{0.09}$ french$^{0.09}$
28	like$^{0.67}$ just$^{0.39}$ better$^{0.32}$ bad$^{0.27}$ think$^{0.26}$ don't$^{0.19}$ say$^{0.14}$ thought$^{0.12}$ food$^{0.12}$ looked$^{0.09}$
29	place$^{1.35}$ it's$^{0.85}$ eat$^{0.45}$ worth$^{0.19}$ you're$^{0.14}$ small$^{0.12}$ can't$^{0.12}$ wait$^{0.10}$ expect$^{0.08}$ crowded$^{0.08}$

Tabela 13: Tópicos estimados com o modelo **LDA** e **50 tópicos** no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	restaurant$^{1.02}$ recommend$^{0.78}$ food$^{0.50}$ highly$^{0.47}$ excellent$^{0.23}$ service$^{0.22}$ delicious$^{0.18}$ recommended$^{0.13}$ priced$^{0.11}$ reasonably$^{0.07}$
1	pizza$^{1.35}$ cheese$^{0.43}$ good$^{0.18}$ order$^{0.15}$ crust$^{0.15}$ pizzas$^{0.11}$ like$^{0.09}$ toppings$^{0.08}$ place$^{0.07}$ pie$^{0.06}$
2	excellent$^{1.20}$ service$^{1.19}$ food$^{0.80}$ wonderful$^{0.46}$ return$^{0.26}$ definitely$^{0.21}$ atmosphere$^{0.16}$ absolutely$^{0.12}$ amazing$^{0.12}$ superb$^{0.08}$
3	steak$^{0.71}$ wife$^{0.33}$ just$^{0.26}$ ordered$^{0.22}$ dinner$^{0.18}$ cooked$^{0.16}$ rib$^{0.15}$ steaks$^{0.12}$ lamb$^{0.12}$ prime$^{0.12}$
4	it's$^{1.52}$ place$^{0.70}$ you're$^{0.26}$ don't$^{0.24}$ people$^{0.15}$ sure$^{0.12}$ isn't$^{0.11}$ i'm$^{0.11}$ know$^{0.10}$ want$^{0.10}$
5	2$^{0.42}$ party$^{0.33}$ people$^{0.30}$ 3$^{0.28}$ reservations$^{0.22}$ 5$^{0.21}$ 4$^{0.16}$ 1$^{0.15}$ group$^{0.15}$ service$^{0.15}$
6	dessert$^{0.42}$ cream$^{0.40}$ chocolate$^{0.35}$ ice$^{0.29}$ cake$^{0.15}$ hot$^{0.13}$ perfect$^{0.13}$ delicious$^{0.10}$ flavors$^{0.07}$ pie$^{0.07}$
7	place$^{0.35}$ atmosphere$^{0.32}$ music$^{0.26}$ live$^{0.25}$ food$^{0.21}$ enjoy$^{0.20}$ comfortable$^{0.15}$ loud$^{0.15}$ fun$^{0.15}$ make$^{0.11}$
8	italian$^{0.45}$ pasta$^{0.32}$ chicken$^{0.30}$ sauce$^{0.23}$ bread$^{0.18}$ dish$^{0.14}$ dishes$^{0.10}$ greek$^{0.10}$ veal$^{0.09}$ meat$^{0.08}$
9	said$^{0.43}$ told$^{0.34}$ manager$^{0.30}$ waitress$^{0.28}$ asked$^{0.28}$ did$^{0.24}$ left$^{0.14}$ owner$^{0.13}$ came$^{0.13}$ tip$^{0.12}$
10	make$^{0.34}$ way$^{0.30}$ just$^{0.25}$ order$^{0.25}$ sure$^{0.23}$ going$^{0.17}$ door$^{0.15}$ cook$^{0.15}$ open$^{0.14}$ kind$^{0.13}$
11	course$^{0.23}$ main$^{0.20}$ restaurant$^{0.13}$ dessert$^{0.11}$ duck$^{0.11}$ served$^{0.08}$ amazing$^{0.07}$ just$^{0.06}$ charming$^{0.06}$ brought$^{0.05}$
12	prices$^{0.71}$ price$^{0.65}$ quality$^{0.42}$ reasonable$^{0.35}$ portions$^{0.31}$ food$^{0.21}$ small$^{0.18}$ high$^{0.16}$ value$^{0.12}$ restaurants$^{0.12}$
13	ordered$^{0.58}$ better$^{0.32}$ like$^{0.31}$ tasted$^{0.22}$ husband$^{0.19}$ didn't$^{0.17}$ looked$^{0.11}$ bland$^{0.09}$ entree$^{0.09}$ did$^{0.09}$
14	mexican$^{0.37}$ chips$^{0.22}$ chicken$^{0.19}$ salsa$^{0.14}$ beans$^{0.09}$ tacos$^{0.08}$ tasty$^{0.07}$ authentic$^{0.07}$ food$^{0.07}$ taco$^{0.07}$
15	buffet$^{0.57}$ menu$^{0.34}$ variety$^{0.27}$ large$^{0.21}$ selection$^{0.19}$ items$^{0.18}$ offers$^{0.16}$ lunch$^{0.16}$ including$^{0.13}$ offer$^{0.12}$
16	best$^{2.54}$ town$^{0.38}$ try$^{0.32}$ favorite$^{0.18}$ far$^{0.18}$ i've$^{0.17}$ area$^{0.17}$ week$^{0.13}$ vegas$^{0.10}$ can't$^{0.05}$
17	years$^{0.61}$ times$^{0.43}$ i've$^{0.34}$ year$^{0.24}$ eaten$^{0.22}$ we've$^{0.16}$ ago$^{0.16}$ time$^{0.15}$ old$^{0.15}$ gone$^{0.11}$
18	fries$^{0.50}$ sandwich$^{0.50}$ burger$^{0.42}$ cheese$^{0.26}$ chicken$^{0.16}$ burgers$^{0.14}$ bacon$^{0.11}$ menu$^{0.10}$ bun$^{0.08}$ lunch$^{0.08}$
19	food$^{1.19}$ service$^{0.71}$ bad$^{0.57}$ experience$^{0.21}$ horrible$^{0.17}$ cold$^{0.15}$ terrible$^{0.14}$ slow$^{0.13}$ ok$^{0.11}$ rude$^{0.10}$
20	minutes$^{0.57}$ order$^{0.34}$ took$^{0.29}$ waitress$^{0.20}$ got$^{0.19}$ wait$^{0.17}$ arrived$^{0.17}$ waited$^{0.17}$ table$^{0.17}$ finally$^{0.15}$
21	waiter$^{0.51}$ table$^{0.46}$ server$^{0.27}$ water$^{0.20}$ brought$^{0.15}$ seated$^{0.15}$ check$^{0.13}$ ask$^{0.13}$ drink$^{0.12}$ hostess$^{0.11}$
22	restaurant$^{1.51}$ family$^{0.91}$ favorite$^{0.17}$ restaurants$^{0.15}$ friends$^{0.11}$ owned$^{0.10}$ visit$^{0.09}$ wonderful$^{0.09}$ try$^{0.08}$ visiting$^{0.08}$
23	nice$^{0.94}$ little$^{0.45}$ good$^{0.43}$ quite$^{0.42}$ bit$^{0.34}$ outside$^{0.27}$ inside$^{0.25}$ seating$^{0.12}$ patio$^{0.11}$ sit$^{0.09}$
24	shrimp$^{0.49}$ seafood$^{0.46}$ crab$^{0.43}$ fresh$^{0.37}$ fish$^{0.32}$ san$^{0.17}$ lobster$^{0.11}$ salmon$^{0.11}$ bay$^{0.10}$ beach$^{0.08}$
25	bbq$^{0.45}$ sauce$^{0.32}$ pork$^{0.32}$ chicken$^{0.25}$ ribs$^{0.23}$ fried$^{0.22}$ meat$^{0.13}$ sweet$^{0.12}$ beans$^{0.09}$ pulled$^{0.08}$
26	room$^{0.48}$ dining$^{0.32}$ right$^{0.15}$ tables$^{0.14}$ table$^{0.14}$ building$^{0.14}$ huge$^{0.13}$ parking$^{0.11}$ red$^{0.09}$ walls$^{0.09}$
27	bar$^{1.41}$ area$^{0.27}$ drinks$^{0.23}$ beer$^{0.23}$ good$^{0.21}$ pub$^{0.14}$ room$^{0.13}$ drink$^{0.11}$ bartender$^{0.10}$ serve$^{0.09}$
28	served$^{0.20}$ sauce$^{0.16}$ grilled$^{0.14}$ garlic$^{0.10}$ vegetables$^{0.08}$ mushrooms$^{0.07}$ roasted$^{0.07}$ potatoes$^{0.07}$ cheese$^{0.07}$ green$^{0.07}$
29	new$^{1.04}$ try$^{0.31}$ restaurant$^{0.31}$ location$^{0.21}$ visit$^{0.15}$ city$^{0.11}$ york$^{0.11}$ disappointed$^{0.09}$ opened$^{0.08}$ original$^{0.07}$
30	really$^{0.44}$ pretty$^{0.33}$ little$^{0.31}$ got$^{0.27}$ right$^{0.15}$ thing$^{0.13}$ just$^{0.13}$ downtown$^{0.11}$ lunch$^{0.09}$ stopped$^{0.09}$
31	like$^{0.49}$ don't$^{0.24}$ home$^{0.21}$ good$^{0.18}$ wings$^{0.16}$ looking$^{0.15}$ people$^{0.13}$ i'm$^{0.13}$ make$^{0.13}$ think$^{0.12}$
32	experience$^{1.31}$ dining$^{0.83}$ fine$^{0.27}$ beautiful$^{0.26}$ enjoyed$^{0.15}$ truly$^{0.13}$ lovely$^{0.11}$ enjoyable$^{0.07}$ dine$^{0.07}$ delicious$^{0.07}$
33	sushi$^{0.53}$ dishes$^{0.32}$ rice$^{0.25}$ chicken$^{0.22}$ chinese$^{0.21}$ rolls$^{0.16}$ thai$^{0.16}$ roll$^{0.13}$ spicy$^{0.13}$ dish$^{0.11}$
34	cafe$^{0.37}$ local$^{0.26}$ menu$^{0.23}$ restaurants$^{0.14}$ number$^{0.12}$ park$^{0.11}$ area$^{0.09}$ east$^{0.07}$ big$^{0.07}$ coffee$^{0.07}$
35	house$^{0.26}$ located$^{0.16}$ daily$^{0.15}$ american$^{0.15}$ lunch$^{0.15}$ cuisine$^{0.14}$ specials$^{0.12}$ world$^{0.08}$ tea$^{0.08}$ serves$^{0.07}$
36	customers$^{0.23}$ restaurant$^{0.22}$ business$^{0.20}$ owner$^{0.20}$ care$^{0.13}$ management$^{0.13}$ establishment$^{0.11}$ waitresses$^{0.11}$ patrons$^{0.10}$ need$^{0.10}$
37	breakfast$^{0.54}$ diner$^{0.21}$ day$^{0.19}$ hotel$^{0.18}$ sunday$^{0.16}$ lunch$^{0.16}$ eggs$^{0.15}$ brunch$^{0.14}$ sausage$^{0.12}$ bacon$^{0.11}$
38	wine$^{0.92}$ chef$^{0.37}$ list$^{0.34}$ menu$^{0.34}$ french$^{0.13}$ restaurant$^{0.12}$ special$^{0.12}$ bistro$^{0.11}$ glass$^{0.11}$ bottle$^{0.11}$
39	great$^{1.97}$ place$^{1.14}$ love$^{0.67}$ food$^{0.53}$ awesome$^{0.20}$ service$^{0.19}$ wonderful$^{0.07}$ friends$^{0.06}$ loved$^{0.05}$ looking$^{0.03}$
40	salad$^{1.21}$ meal$^{0.44}$ bread$^{0.36}$ soup$^{0.31}$ served$^{0.26}$ menu$^{0.24}$ potato$^{0.15}$ dressing$^{0.14}$ meals$^{0.12}$ homemade$^{0.12}$
41	good$^{2.50}$ food$^{0.70}$ place$^{0.58}$ really$^{0.42}$ busy$^{0.08}$ places$^{0.07}$ real$^{0.05}$ town$^{0.04}$ plenty$^{0.04}$ usually$^{0.02}$
42	great$^{2.76}$ food$^{0.72}$ atmosphere$^{0.62}$ service$^{0.38}$ cozy$^{0.14}$ fantastic$^{0.13}$ wonderful$^{0.11}$ friendly$^{0.06}$ kids$^{0.05}$ restaurant$^{0.05}$
43	dinner$^{0.63}$ night$^{0.46}$ went$^{0.41}$ friends$^{0.20}$ friend$^{0.19}$ evening$^{0.16}$ special$^{0.15}$ going$^{0.14}$ husband$^{0.14}$ decided$^{0.14}$
44	time$^{0.82}$ worth$^{0.62}$ wait$^{0.52}$ long$^{0.34}$ visit$^{0.33}$ trip$^{0.18}$ money$^{0.12}$ second$^{0.11}$ return$^{0.10}$ drive$^{0.10}$
45	just$^{0.49}$ know$^{0.39}$ don't$^{0.34}$ want$^{0.30}$ say$^{0.27}$ think$^{0.27}$ like$^{0.25}$ didn't$^{0.20}$ thought$^{0.18}$ let$^{0.09}$
46	staff$^{1.32}$ friendly$^{1.13}$ food$^{0.31}$ wait$^{0.26}$ attentive$^{0.26}$ pleasant$^{0.23}$ service$^{0.14}$ helpful$^{0.12}$ enjoyed$^{0.10}$ atmosphere$^{0.09}$
47	got$^{0.29}$ really$^{0.25}$ came$^{0.23}$ didn't$^{0.19}$ little$^{0.12}$ just$^{0.11}$ j$^{0.10}$ there's$^{0.10}$ wasn't$^{0.10}$ said$^{0.09}$
48	like$^{0.52}$ beef$^{0.49}$ hot$^{0.39}$ just$^{0.31}$ style$^{0.28}$ chicago$^{0.25}$ italian$^{0.17}$ places$^{0.17}$ place$^{0.15}$ sandwiches$^{0.13}$
49	eat$^{1.29}$ food$^{0.50}$ time$^{0.47}$ meal$^{0.33}$ eating$^{0.20}$ ate$^{0.15}$ person$^{0.14}$ day$^{0.08}$ say$^{0.08}$ going$^{0.08}$

Tabela 14: Tópicos estimados com o modelo de **temas** e **10 tópicos** no conjunto de dados we8there utlizando as avaliações do aspecto *comida*. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	shrimp$^{0.07}$ sauce$^{0.06}$ salad$^{0.04}$ garlic$^{0.04}$ chicken$^{0.04}$ crab$^{0.03}$ fried$^{0.03}$ grilled$^{0.03}$ potatoes$^{0.03}$ cheese$^{0.03}$
1	sushi$^{0.21}$ chinese$^{0.10}$ buffet$^{0.09}$ mexican$^{0.08}$ rice$^{0.07}$ thai$^{0.07}$ dishes$^{0.06}$ shrimp$^{0.04}$ japanese$^{0.04}$ salsa$^{0.04}$
2	minutes$^{0.06}$ asked$^{0.06}$ manager$^{0.06}$ waitress$^{0.05}$ did$^{0.04}$ service$^{0.04}$ told$^{0.04}$ came$^{0.03}$ table$^{0.03}$ said$^{0.03}$
3	seafood$^{0.04}$ bar$^{0.04}$ menu$^{0.03}$ located$^{0.03}$ friendly$^{0.02}$ dining$^{0.02}$ spot$^{0.02}$ offers$^{0.02}$ room$^{0.02}$ casual$^{0.02}$
4	burger$^{0.09}$ fries$^{0.07}$ got$^{0.06}$ sandwich$^{0.05}$ came$^{0.05}$ really$^{0.05}$ j$^{0.04}$ pretty$^{0.03}$ cheese$^{0.03}$ didn't$^{0.03}$
5	great$^{0.20}$ food$^{0.13}$ service$^{0.11}$ excellent$^{0.07}$ wonderful$^{0.06}$ friendly$^{0.06}$ place$^{0.05}$ atmosphere$^{0.05}$ staff$^{0.04}$ best$^{0.03}$
6	breakfast$^{0.24}$ coffee$^{0.10}$ eggs$^{0.09}$ cream$^{0.08}$ cafe$^{0.07}$ chocolate$^{0.06}$ toast$^{0.06}$ ice$^{0.05}$ brunch$^{0.05}$ diner$^{0.04}$
7	wine$^{0.11}$ chef$^{0.05}$ experience$^{0.04}$ dining$^{0.03}$ restaurant$^{0.03}$ list$^{0.03}$ duck$^{0.02}$ presentation$^{0.02}$ wines$^{0.02}$ evening$^{0.02}$
8	food$^{0.05}$ place$^{0.05}$ bad$^{0.04}$ good$^{0.03}$ like$^{0.03}$ don't$^{0.03}$ eat$^{0.02}$ it's$^{0.02}$ just$^{0.02}$ pretty$^{0.02}$
9	pizza$^{0.50}$ bbq$^{0.14}$ cheese$^{0.06}$ chicago$^{0.06}$ ribs$^{0.06}$ wings$^{0.05}$ crust$^{0.05}$ pork$^{0.04}$ pizzas$^{0.04}$ barbecue$^{0.03}$

Tabela 15: Tópicos estimados com o modelo de **temas** e **10 tópicos** no conjunto de dados we8there utlizando as avaliações do aspecto *serviço*. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	sauce$^{0.05}$ salad$^{0.05}$ crab$^{0.04}$ shrimp$^{0.04}$ garlic$^{0.03}$ chocolate$^{0.03}$ ordered$^{0.03}$ chicken$^{0.02}$ grilled$^{0.02}$ bread$^{0.02}$
1	mexican$^{0.21}$ chips$^{0.10}$ salsa$^{0.10}$ tacos$^{0.06}$ taco$^{0.06}$ burrito$^{0.04}$ cheese$^{0.04}$ cuban$^{0.03}$ chile$^{0.03}$ tortilla$^{0.03}$
2	minutes$^{0.06}$ asked$^{0.06}$ manager$^{0.05}$ waitress$^{0.05}$ told$^{0.04}$ service$^{0.04}$ did$^{0.03}$ came$^{0.03}$ took$^{0.03}$ waiter$^{0.03}$
3	breakfast$^{0.22}$ cafe$^{0.07}$ coffee$^{0.06}$ eggs$^{0.06}$ brunch$^{0.05}$ chocolate$^{0.04}$ cream$^{0.04}$ sunday$^{0.03}$ toast$^{0.03}$ ice$^{0.03}$
4	burger$^{0.08}$ fries$^{0.06}$ came$^{0.06}$ got$^{0.06}$ sandwich$^{0.06}$ really$^{0.05}$ j$^{0.04}$ cheese$^{0.04}$ pretty$^{0.03}$ little$^{0.03}$
5	great$^{0.18}$ food$^{0.12}$ service$^{0.10}$ excellent$^{0.06}$ friendly$^{0.06}$ wonderful$^{0.06}$ place$^{0.06}$ staff$^{0.05}$ atmosphere$^{0.04}$ best$^{0.03}$
6	sushi$^{0.26}$ buffet$^{0.15}$ chinese$^{0.11}$ thai$^{0.08}$ dishes$^{0.06}$ rice$^{0.05}$ indian$^{0.05}$ japanese$^{0.05}$ shrimp$^{0.05}$ roll$^{0.04}$
7	wine$^{0.09}$ dining$^{0.04}$ menu$^{0.04}$ seafood$^{0.04}$ list$^{0.03}$ restaurant$^{0.03}$ staff$^{0.02}$ located$^{0.02}$ bar$^{0.02}$ room$^{0.02}$
8	place$^{0.06}$ food$^{0.06}$ good$^{0.04}$ it's$^{0.03}$ like$^{0.03}$ just$^{0.03}$ bad$^{0.03}$ don't$^{0.02}$ eat$^{0.02}$ really$^{0.02}$
9	pizza$^{0.53}$ bbq$^{0.14}$ chicago$^{0.07}$ ribs$^{0.06}$ wings$^{0.06}$ crust$^{0.05}$ cheese$^{0.04}$ pork$^{0.04}$ pizzas$^{0.04}$ sauce$^{0.03}$

Tabela 16: Tópicos estimados com o modelo de **temas** e **15 tópicos** no conjunto de dados we8there utlizando as avaliações do aspecto *comida*. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	breakfast$^{0.27}$ eggs$^{0.10}$ cream$^{0.10}$ coffee$^{0.09}$ chocolate$^{0.08}$ ice$^{0.07}$ cafe$^{0.06}$ toast$^{0.05}$ brunch$^{0.05}$ shop$^{0.04}$
1	pizza$^{0.56}$ chicago$^{0.08}$ crust$^{0.06}$ cheese$^{0.05}$ pizzas$^{0.05}$ italian$^{0.04}$ pepperoni$^{0.04}$ pretty$^{0.03}$ style$^{0.03}$ shop$^{0.03}$
2	food$^{0.04}$ bland$^{0.03}$ ordered$^{0.03}$ overcooked$^{0.02}$ salad$^{0.02}$ worst$^{0.02}$ tasted$^{0.02}$ tasteless$^{0.02}$ shrimp$^{0.02}$ frozen$^{0.02}$
3	bar$^{0.13}$ music$^{0.07}$ beer$^{0.07}$ pub$^{0.06}$ irish$^{0.06}$ great$^{0.05}$ place$^{0.03}$ beers$^{0.03}$ german$^{0.02}$ patio$^{0.02}$
4	best$^{0.07}$ years$^{0.05}$ crab$^{0.02}$ year$^{0.02}$ family$^{0.02}$ steak$^{0.02}$ visit$^{0.02}$ time$^{0.02}$ stop$^{0.02}$ live$^{0.02}$
5	minutes$^{0.11}$ waitress$^{0.07}$ asked$^{0.06}$ manager$^{0.04}$ waiter$^{0.04}$ told$^{0.04}$ came$^{0.04}$ took$^{0.04}$ table$^{0.04}$ seated$^{0.03}$
6	place$^{0.05}$ it's$^{0.04}$ good$^{0.03}$ just$^{0.03}$ know$^{0.02}$ like$^{0.02}$ don't$^{0.02}$ eat$^{0.02}$ didn't$^{0.02}$ food$^{0.02}$
7	garlic$^{0.04}$ sauce$^{0.04}$ shrimp$^{0.04}$ chicken$^{0.03}$ crab$^{0.03}$ served$^{0.03}$ salad$^{0.03}$ grilled$^{0.03}$ cheese$^{0.03}$ potatoes$^{0.02}$
8	menu$^{0.05}$ seafood$^{0.04}$ located$^{0.03}$ offers$^{0.02}$ casual$^{0.02}$ dining$^{0.02}$ restaurant$^{0.02}$ daily$^{0.02}$ selections$^{0.02}$ fish$^{0.02}$
9	great$^{0.15}$ food$^{0.10}$ service$^{0.08}$ excellent$^{0.07}$ friendly$^{0.06}$ wonderful$^{0.06}$ atmosphere$^{0.06}$ staff$^{0.04}$ place$^{0.04}$ good$^{0.03}$
10	sushi$^{0.29}$ chinese$^{0.09}$ buffet$^{0.09}$ thai$^{0.09}$ rice$^{0.08}$ dishes$^{0.06}$ japanese$^{0.05}$ indian$^{0.05}$ roll$^{0.04}$ shrimp$^{0.04}$
11	owner$^{0.04}$ manager$^{0.04}$ restaurant$^{0.04}$ customer$^{0.03}$ rude$^{0.03}$ card$^{0.03}$ service$^{0.03}$ staff$^{0.03}$ credit$^{0.03}$ customers$^{0.03}$
12	burger$^{0.08}$ fries$^{0.06}$ sandwich$^{0.05}$ got$^{0.04}$ j$^{0.04}$ came$^{0.04}$ really$^{0.04}$ cheese$^{0.03}$ didn't$^{0.02}$ wasn't$^{0.02}$
13	wine$^{0.10}$ chef$^{0.04}$ list$^{0.03}$ dining$^{0.02}$ duck$^{0.02}$ experience$^{0.02}$ wines$^{0.02}$ bottle$^{0.02}$ evening$^{0.02}$ course$^{0.02}$
14	bbq$^{0.15}$ mexican$^{0.12}$ pork$^{0.07}$ ribs$^{0.06}$ beans$^{0.06}$ salsa$^{0.05}$ tacos$^{0.04}$ wings$^{0.04}$ slaw$^{0.04}$ taco$^{0.03}$

Tabela 17: Tópicos estimados com o modelo de **temas** e **15 tópicos** no conjunto de dados we8there utlizando as avaliações do aspecto *serviço*. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	breakfast$^{0.32}$ coffee$^{0.14}$ eggs$^{0.11}$ chocolate$^{0.09}$ cream$^{0.08}$ toast$^{0.06}$ brunch$^{0.06}$ cafe$^{0.05}$ diner$^{0.05}$ ice$^{0.05}$
1	pizza$^{0.71}$ cheese$^{0.09}$ chicago$^{0.08}$ crust$^{0.07}$ pizzas$^{0.06}$ deli$^{0.05}$ italian$^{0.05}$ pepperoni$^{0.04}$ shop$^{0.03}$ sandwiches$^{0.03}$
2	ordered$^{0.06}$ salad$^{0.04}$ food$^{0.04}$ bland$^{0.04}$ overcooked$^{0.03}$ good$^{0.03}$ husband$^{0.03}$ steak$^{0.03}$ service$^{0.02}$ tasted$^{0.02}$
3	great$^{0.18}$ food$^{0.15}$ service$^{0.10}$ friendly$^{0.08}$ excellent$^{0.07}$ wonderful$^{0.07}$ atmosphere$^{0.07}$ staff$^{0.06}$ place$^{0.05}$ recommend$^{0.04}$
4	years$^{0.07}$ best$^{0.06}$ new$^{0.03}$ restaurant$^{0.03}$ family$^{0.03}$ year$^{0.03}$ live$^{0.03}$ crab$^{0.02}$ stop$^{0.02}$ wish$^{0.02}$
5	minutes$^{0.12}$ waitress$^{0.07}$ asked$^{0.07}$ table$^{0.05}$ told$^{0.05}$ waiter$^{0.04}$ came$^{0.04}$ seated$^{0.04}$ waited$^{0.04}$ took$^{0.04}$
6	it's$^{0.06}$ place$^{0.06}$ don't$^{0.04}$ just$^{0.03}$ know$^{0.03}$ i'm$^{0.03}$ want$^{0.03}$ eat$^{0.03}$ like$^{0.03}$ didn't$^{0.03}$
7	sauce$^{0.06}$ garlic$^{0.05}$ chicken$^{0.04}$ shrimp$^{0.04}$ salad$^{0.04}$ grilled$^{0.04}$ cheese$^{0.03}$ served$^{0.03}$ crab$^{0.03}$ fresh$^{0.03}$
8	bar$^{0.19}$ great$^{0.08}$ music$^{0.08}$ pub$^{0.06}$ beer$^{0.06}$ irish$^{0.05}$ place$^{0.04}$ vegas$^{0.04}$ beers$^{0.03}$ german$^{0.03}$
9	wine$^{0.12}$ experience$^{0.05}$ chef$^{0.04}$ list$^{0.03}$ duck$^{0.03}$ presentation$^{0.02}$ bistro$^{0.02}$ amazing$^{0.02}$ course$^{0.02}$ dessert$^{0.02}$
10	sushi$^{0.33}$ buffet$^{0.18}$ chinese$^{0.13}$ rice$^{0.09}$ thai$^{0.09}$ dishes$^{0.06}$ indian$^{0.06}$ japanese$^{0.06}$ roll$^{0.05}$ chicken$^{0.04}$
11	manager$^{0.05}$ rude$^{0.04}$ owner$^{0.04}$ service$^{0.04}$ customer$^{0.04}$ experience$^{0.04}$ management$^{0.03}$ restaurant$^{0.03}$ card$^{0.03}$ food$^{0.03}$
12	burger$^{0.08}$ fries$^{0.07}$ sandwich$^{0.06}$ got$^{0.05}$ came$^{0.05}$ j$^{0.04}$ really$^{0.04}$ cheese$^{0.04}$ wasn't$^{0.03}$ didn't$^{0.03}$
13	menu$^{0.05}$ seafood$^{0.04}$ located$^{0.03}$ offers$^{0.03}$ dining$^{0.02}$ available$^{0.02}$ casual$^{0.02}$ cuisine$^{0.02}$ wine$^{0.02}$ selections$^{0.02}$
14	bbq$^{0.17}$ mexican$^{0.14}$ pork$^{0.09}$ ribs$^{0.08}$ beans$^{0.07}$ salsa$^{0.06}$ wings$^{0.06}$ chips$^{0.05}$ tacos$^{0.05}$ sauce$^{0.05}$

Tabela 18: Tópicos estimados com o modelo de **temas** e **30 tópicos** no conjunto de dados we8there utlizando as avaliações do aspecto *comida*. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	don't$^{0.10}$ it's$^{0.09}$ place$^{0.07}$ i'm$^{0.06}$ know$^{0.06}$ just$^{0.05}$ want$^{0.05}$ like$^{0.05}$ think$^{0.05}$ eat$^{0.04}$
1	breakfast$^{0.61}$ eggs$^{0.21}$ toast$^{0.13}$ coffee$^{0.13}$ bacon$^{0.13}$ brunch$^{0.13}$ diner$^{0.10}$ pancakes$^{0.09}$ buffet$^{0.09}$ omelet$^{0.08}$
2	shrimp$^{0.32}$ crab$^{0.28}$ seafood$^{0.24}$ lobster$^{0.13}$ steak$^{0.10}$ filet$^{0.10}$ rib$^{0.08}$ prime$^{0.07}$ fish$^{0.07}$ cakes$^{0.06}$
3	sauce$^{0.12}$ salad$^{0.09}$ garlic$^{0.08}$ cheese$^{0.07}$ pasta$^{0.06}$ chicken$^{0.06}$ bread$^{0.06}$ served$^{0.05}$ italian$^{0.05}$ potatoes$^{0.05}$
4	best$^{0.20}$ years$^{0.10}$ i've$^{0.07}$ live$^{0.06}$ san$^{0.06}$ new$^{0.04}$ vegas$^{0.04}$ try$^{0.04}$ love$^{0.04}$ las$^{0.04}$
5	minutes$^{0.28}$ waited$^{0.09}$ seated$^{0.09}$ arrived$^{0.09}$ table$^{0.08}$ wait$^{0.08}$ took$^{0.07}$ order$^{0.06}$ waiter$^{0.06}$ waitress$^{0.05}$
6	great$^{0.50}$ food$^{0.18}$ place$^{0.17}$ service$^{0.11}$ love$^{0.10}$ friendly$^{0.09}$ wonderful$^{0.07}$ it's$^{0.06}$ good$^{0.05}$ eat$^{0.04}$
7	burger$^{0.45}$ sandwich$^{0.38}$ fries$^{0.29}$ cheese$^{0.17}$ burgers$^{0.10}$ bacon$^{0.09}$ onion$^{0.06}$ bun$^{0.06}$ lettuce$^{0.05}$ sandwiches$^{0.05}$
8	manager$^{0.15}$ told$^{0.14}$ said$^{0.13}$ asked$^{0.11}$ rude$^{0.10}$ owner$^{0.07}$ did$^{0.07}$ waitress$^{0.05}$ charged$^{0.04}$ came$^{0.04}$
9	wings$^{0.46}$ chicken$^{0.20}$ fries$^{0.18}$ buffalo$^{0.17}$ wing$^{0.12}$ joe's$^{0.11}$ jimmy's$^{0.10}$ pearl's$^{0.07}$ fried$^{0.07}$ hot$^{0.05}$
10	years$^{0.15}$ year$^{0.12}$ family$^{0.09}$ children$^{0.07}$ wife$^{0.06}$ ago$^{0.05}$ 4$^{0.05}$ kids$^{0.05}$ cards$^{0.05}$ 2$^{0.05}$
11	water$^{0.12}$ glass$^{0.09}$ salad$^{0.08}$ waitress$^{0.08}$ asked$^{0.07}$ bread$^{0.06}$ glasses$^{0.05}$ tea$^{0.05}$ brought$^{0.04}$ table$^{0.04}$
12	pizza$^{1.75}$ cheese$^{0.17}$ crust$^{0.17}$ pizzas$^{0.12}$ pepperoni$^{0.10}$ chicago$^{0.09}$ toppings$^{0.06}$ pizzeria$^{0.05}$ pie$^{0.05}$ order$^{0.05}$
13	got$^{0.09}$ really$^{0.09}$ came$^{0.08}$ j$^{0.08}$ didn't$^{0.06}$ wasn't$^{0.05}$ pretty$^{0.05}$ there's$^{0.05}$ little$^{0.05}$ girlfriend$^{0.04}$
14	good$^{0.22}$ food$^{0.11}$ service$^{0.06}$ price$^{0.05}$ place$^{0.05}$ nice$^{0.05}$ little$^{0.05}$ priced$^{0.04}$ atmosphere$^{0.04}$ overall$^{0.04}$
15	bbq$^{0.43}$ pork$^{0.21}$ ribs$^{0.19}$ southern$^{0.09}$ pulled$^{0.09}$ sauce$^{0.09}$ slaw$^{0.08}$ beans$^{0.08}$ barbecue$^{0.08}$ catfish$^{0.07}$
16	sushi$^{1.14}$ roll$^{0.21}$ japanese$^{0.17}$ tuna$^{0.13}$ rolls$^{0.12}$ sashimi$^{0.09}$ fresh$^{0.08}$ chef$^{0.08}$ tempura$^{0.07}$ sake$^{0.07}$
17	mexican$^{0.33}$ salsa$^{0.15}$ indian$^{0.14}$ chips$^{0.13}$ tacos$^{0.11}$ taco$^{0.11}$ burrito$^{0.08}$ beans$^{0.07}$ tortilla$^{0.06}$ chicken$^{0.06}$
18	bar$^{0.39}$ music$^{0.17}$ pub$^{0.13}$ beer$^{0.12}$ great$^{0.11}$ irish$^{0.11}$ beers$^{0.08}$ playing$^{0.06}$ live$^{0.05}$ night$^{0.05}$
19	dining$^{0.08}$ room$^{0.08}$ beautiful$^{0.06}$ patio$^{0.05}$ seating$^{0.04}$ building$^{0.04}$ main$^{0.04}$ walls$^{0.04}$ parking$^{0.03}$ street$^{0.03}$
20	chocolate$^{0.54}$ ice$^{0.29}$ cream$^{0.25}$ dessert$^{0.22}$ cake$^{0.14}$ creme$^{0.09}$ coffee$^{0.07}$ cheesecake$^{0.06}$ brulee$^{0.05}$ pie$^{0.05}$
21	food$^{0.15}$ atmosphere$^{0.12}$ excellent$^{0.12}$ service$^{0.12}$ experience$^{0.11}$ staff$^{0.10}$ wonderful$^{0.09}$ friendly$^{0.08}$ highly$^{0.08}$ recommend$^{0.07}$
22	food$^{0.13}$ bad$^{0.11}$ worst$^{0.06}$ terrible$^{0.06}$ ordered$^{0.06}$ bland$^{0.06}$ horrible$^{0.05}$ better$^{0.05}$ ok$^{0.05}$ tasted$^{0.04}$
23	buffet$^{0.31}$ chinese$^{0.24}$ thai$^{0.18}$ dishes$^{0.13}$ rice$^{0.10}$ chicken$^{0.10}$ shrimp$^{0.06}$ curry$^{0.06}$ china$^{0.05}$ sushi$^{0.04}$
24	deli$^{0.13}$ beef$^{0.11}$ sandwhich$^{0.10}$ italian$^{0.09}$ chicago$^{0.09}$ fries$^{0.08}$ sandwich$^{0.07}$ peoria$^{0.07}$ shop$^{0.07}$ downtown$^{0.07}$
25	husband$^{0.08}$ birthday$^{0.07}$ went$^{0.06}$ ordered$^{0.06}$ wife$^{0.06}$ night$^{0.06}$ friends$^{0.05}$ friend$^{0.05}$ dinner$^{0.04}$ wine$^{0.04}$
26	wine$^{0.23}$ chef$^{0.13}$ list$^{0.09}$ duck$^{0.07}$ experience$^{0.06}$ cuisine$^{0.05}$ dishes$^{0.04}$ wines$^{0.04}$ bistro$^{0.04}$ french$^{0.04}$
27	menu$^{0.16}$ seafood$^{0.13}$ variety$^{0.06}$ offers$^{0.06}$ specials$^{0.06}$ wine$^{0.05}$ entrees$^{0.05}$ list$^{0.05}$ large$^{0.05}$ friendly$^{0.05}$
28	restaurant$^{0.07}$ staff$^{0.06}$ review$^{0.06}$ management$^{0.05}$ visit$^{0.05}$ needs$^{0.04}$ patrons$^{0.04}$ experience$^{0.04}$ service$^{0.04}$ establishment$^{0.03}$
29	cafe$^{0.09}$ it's$^{0.05}$ saturday$^{0.04}$ friday$^{0.04}$ san$^{0.04}$ you'll$^{0.04}$ there's$^{0.03}$ monday$^{0.03}$ priced$^{0.03}$ you're$^{0.03}$

Tabela 19: Tópicos estimados com o modelo de **temas** e **30 tópicos** no conjunto de dados we8there utlizando as avaliações do aspecto *serviço*. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	sandwich$^{0.61}$ sandwiches$^{0.16}$ deli$^{0.16}$ cheese$^{0.10}$ turkey$^{0.10}$ salad$^{0.10}$ bread$^{0.09}$ cafe$^{0.08}$ provolone$^{0.05}$ beef$^{0.05}$
1	breakfast$^{0.70}$ eggs$^{0.25}$ toast$^{0.14}$ brunch$^{0.13}$ bacon$^{0.11}$ diner$^{0.11}$ pancakes$^{0.09}$ coffee$^{0.09}$ sausage$^{0.08}$ omelet$^{0.08}$
2	steak$^{0.28}$ rib$^{0.16}$ prime$^{0.13}$ potatoes$^{0.09}$ filet$^{0.09}$ ordered$^{0.08}$ lobster$^{0.07}$ steaks$^{0.07}$ medium$^{0.07}$ mashed$^{0.07}$
3	served$^{0.06}$ garlic$^{0.05}$ sauteed$^{0.05}$ rich$^{0.04}$ sauce$^{0.04}$ mushrooms$^{0.04}$ potatoes$^{0.04}$ roasted$^{0.04}$ sunday$^{0.03}$ include$^{0.03}$
4	san$^{0.18}$ cafe$^{0.15}$ francisco$^{0.10}$ california$^{0.06}$ beach$^{0.05}$ bay$^{0.05}$ local$^{0.05}$ west$^{0.05}$ com$^{0.04}$ best$^{0.04}$
5	minutes$^{0.25}$ seated$^{0.09}$ waited$^{0.09}$ took$^{0.08}$ table$^{0.08}$ wait$^{0.07}$ arrived$^{0.07}$ waitress$^{0.07}$ order$^{0.06}$ hour$^{0.06}$
6	great$^{0.49}$ food$^{0.23}$ place$^{0.18}$ service$^{0.14}$ love$^{0.10}$ friendly$^{0.09}$ best$^{0.07}$ atmosphere$^{0.05}$ wonderful$^{0.05}$ prices$^{0.05}$
7	burger$^{0.58}$ fries$^{0.36}$ sandwich$^{0.20}$ burgers$^{0.11}$ cheese$^{0.10}$ bacon$^{0.09}$ bun$^{0.07}$ hamburger$^{0.06}$ onion$^{0.05}$ ketchup$^{0.04}$
8	food$^{0.18}$ bad$^{0.08}$ terrible$^{0.07}$ service$^{0.07}$ horrible$^{0.06}$ worst$^{0.06}$ bland$^{0.05}$ experience$^{0.05}$ awful$^{0.04}$ cold$^{0.04}$
9	wings$^{0.17}$ fries$^{0.17}$ sandwhich$^{0.13}$ chicago$^{0.12}$ italian$^{0.10}$ beef$^{0.09}$ chicken$^{0.08}$ sandwhiches$^{0.08}$ hot$^{0.07}$ peoria$^{0.07}$
10	years$^{0.22}$ year$^{0.09}$ ago$^{0.07}$ live$^{0.05}$ times$^{0.05}$ vegas$^{0.05}$ new$^{0.05}$ las$^{0.04}$ love$^{0.04}$ wish$^{0.04}$
11	manager$^{0.14}$ said$^{0.12}$ asked$^{0.12}$ told$^{0.11}$ owner$^{0.07}$ rude$^{0.06}$ waitress$^{0.06}$ did$^{0.05}$ service$^{0.05}$ card$^{0.05}$
12	pizza$^{1.66}$ crust$^{0.17}$ cheese$^{0.14}$ pizzas$^{0.13}$ chicago$^{0.09}$ pepperoni$^{0.09}$ pizzeria$^{0.05}$ toppings$^{0.04}$ style$^{0.04}$ best$^{0.04}$
13	got$^{0.11}$ came$^{0.09}$ really$^{0.08}$ j$^{0.08}$ didn't$^{0.07}$ pretty$^{0.06}$ wasn't$^{0.06}$ little$^{0.04}$ there's$^{0.04}$ girlfriend$^{0.04}$
14	buffet$^{0.18}$ seafood$^{0.16}$ variety$^{0.10}$ menu$^{0.09}$ dishes$^{0.08}$ offers$^{0.07}$ selections$^{0.06}$ wide$^{0.06}$ lunch$^{0.05}$ items$^{0.05}$
15	bbq$^{0.43}$ pork$^{0.23}$ ribs$^{0.20}$ pulled$^{0.09}$ slaw$^{0.09}$ sauce$^{0.09}$ barbecue$^{0.08}$ beans$^{0.08}$ smoked$^{0.07}$ catfish$^{0.07}$
16	sushi$^{0.60}$ chinese$^{0.23}$ thai$^{0.18}$ rice$^{0.13}$ japanese$^{0.11}$ roll$^{0.11}$ rolls$^{0.09}$ dishes$^{0.07}$ shrimp$^{0.07}$ spicy$^{0.05}$
17	mexican$^{0.33}$ salsa$^{0.16}$ indian$^{0.15}$ chips$^{0.13}$ tacos$^{0.11}$ taco$^{0.10}$ beans$^{0.08}$ burrito$^{0.08}$ rice$^{0.06}$ chicken$^{0.06}$
18	bar$^{0.34}$ music$^{0.14}$ pub$^{0.13}$ beer$^{0.13}$ great$^{0.12}$ irish$^{0.10}$ beers$^{0.08}$ german$^{0.07}$ live$^{0.06}$ bartender$^{0.06}$
19	room$^{0.08}$ dining$^{0.06}$ located$^{0.05}$ beautiful$^{0.04}$ parking$^{0.04}$ building$^{0.04}$ old$^{0.04}$ decor$^{0.03}$ street$^{0.03}$ patio$^{0.03}$
20	chocolate$^{0.49}$ cream$^{0.26}$ ice$^{0.26}$ dessert$^{0.16}$ cake$^{0.13}$ coffee$^{0.06}$ shop$^{0.06}$ truffles$^{0.05}$ hot$^{0.05}$ pie$^{0.05}$
21	food$^{0.14}$ excellent$^{0.14}$ service$^{0.12}$ staff$^{0.12}$ atmosphere$^{0.11}$ experience$^{0.11}$ wonderful$^{0.09}$ friendly$^{0.09}$ restaurant$^{0.06}$ highly$^{0.06}$
22	salad$^{0.13}$ pasta$^{0.13}$ sauce$^{0.10}$ italian$^{0.09}$ bread$^{0.08}$ garlic$^{0.07}$ cheese$^{0.06}$ tomato$^{0.06}$ chicken$^{0.05}$ ravioli$^{0.05}$
23	shrimp$^{0.35}$ crab$^{0.34}$ seafood$^{0.23}$ lobster$^{0.10}$ clam$^{0.07}$ cakes$^{0.07}$ oysters$^{0.07}$ chowder$^{0.07}$ scallops$^{0.06}$ fish$^{0.06}$
24	good$^{0.23}$ food$^{0.19}$ place$^{0.11}$ service$^{0.11}$ price$^{0.07}$ pretty$^{0.06}$ really$^{0.05}$ prices$^{0.05}$ priced$^{0.05}$ portions$^{0.05}$
25	husband$^{0.14}$ went$^{0.07}$ dinner$^{0.06}$ birthday$^{0.06}$ didn't$^{0.05}$ friend$^{0.04}$ did$^{0.04}$ time$^{0.04}$ ordered$^{0.04}$ night$^{0.04}$
26	joe's$^{0.07}$ vienna$^{0.06}$ charming$^{0.05}$ dessert$^{0.05}$ kauai$^{0.05}$ number$^{0.04}$ coffee$^{0.04}$ feast$^{0.04}$ young$^{0.04}$ mai$^{0.04}$
27	wine$^{0.31}$ chef$^{0.12}$ list$^{0.10}$ dishes$^{0.06}$ menu$^{0.05}$ bottle$^{0.05}$ duck$^{0.05}$ wines$^{0.05}$ cuisine$^{0.03}$ french$^{0.03}$
28	restaurant$^{0.09}$ staff$^{0.05}$ customers$^{0.04}$ management$^{0.04}$ patrons$^{0.04}$ review$^{0.04}$ experience$^{0.03}$ customer$^{0.03}$ servers$^{0.03}$ needs$^{0.03}$
29	it's$^{0.12}$ don't$^{0.09}$ place$^{0.07}$ i'm$^{0.07}$ just$^{0.07}$ eat$^{0.07}$ like$^{0.06}$ know$^{0.06}$ want$^{0.05}$ you're$^{0.05}$

Tabela 20: Tópicos estimados com o modelo **Nota-Aspecto** e **10 tópicos** (5 avaliativos e nenhum não avaliativo) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	food$^{0.49}$ people$^{0.10}$ it's$^{0.08}$ cheese$^{0.07}$ dinner$^{0.07}$ area$^{0.06}$ dining$^{0.05}$ lunch$^{0.05}$ wait$^{0.05}$ bad$^{0.04}$
comida+	food$^{0.40}$ it's$^{0.11}$ cheese$^{0.09}$ area$^{0.08}$ people$^{0.08}$ dining$^{0.07}$ wait$^{0.06}$ really$^{0.06}$ lunch$^{0.05}$ sandwich$^{0.05}$
serviço-	got$^{0.13}$ came$^{0.09}$ minutes$^{0.07}$ come$^{0.06}$ great$^{0.06}$ room$^{0.05}$ order$^{0.05}$ know$^{0.05}$ times$^{0.05}$ did$^{0.04}$
serviço+	great$^{0.19}$ staff$^{0.09}$ family$^{0.07}$ got$^{0.07}$ food$^{0.06}$ best$^{0.05}$ times$^{0.05}$ restaurant$^{0.05}$ came$^{0.04}$ place$^{0.04}$
valor-	good$^{0.26}$ like$^{0.16}$ experience$^{0.09}$ chicken$^{0.09}$ night$^{0.07}$ bar$^{0.06}$ say$^{0.06}$ eat$^{0.06}$ place$^{0.06}$ fries$^{0.05}$
valor+	good$^{0.24}$ like$^{0.14}$ chicken$^{0.11}$ eat$^{0.10}$ experience$^{0.08}$ order$^{0.07}$ home$^{0.07}$ i've$^{0.06}$ right$^{0.06}$ night$^{0.05}$
ambiente-	restaurant$^{0.22}$ service$^{0.16}$ time$^{0.13}$ meal$^{0.12}$ small$^{0.07}$ served$^{0.05}$ just$^{0.05}$ pretty$^{0.05}$ make$^{0.05}$ better$^{0.05}$
ambiente+	service$^{0.22}$ restaurant$^{0.21}$ time$^{0.13}$ meal$^{0.11}$ just$^{0.07}$ menu$^{0.06}$ small$^{0.06}$ lot$^{0.05}$ make$^{0.05}$ served$^{0.05}$
experiência-	ordered$^{0.10}$ way$^{0.07}$ did$^{0.06}$ drink$^{0.06}$ think$^{0.05}$ manager$^{0.05}$ drinks$^{0.05}$ waitress$^{0.04}$ just$^{0.04}$ table$^{0.04}$
experiência+	great$^{0.08}$ friendly$^{0.05}$ place$^{0.05}$ wine$^{0.04}$ hot$^{0.04}$ way$^{0.04}$ highly$^{0.04}$ don't$^{0.04}$ think$^{0.04}$ cream$^{0.04}$

Tabela 21: Tópicos estimados com o modelo **Nota-Aspecto** e **15 tópicos** (5 avaliativos e 5 não avaliativos) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	dining$^{0.14}$ bad$^{0.14}$ it's$^{0.14}$ like$^{0.07}$ terrible$^{0.07}$ quality$^{0.07}$ tasted$^{0.06}$ bland$^{0.06}$ buffet$^{0.05}$ told$^{0.05}$
comida+	it's$^{0.23}$ dining$^{0.22}$ good$^{0.08}$ thing$^{0.07}$ pizza$^{0.07}$ fries$^{0.06}$ dishes$^{0.06}$ little$^{0.05}$ menu$^{0.05}$ really$^{0.05}$
serviço-	minutes$^{0.17}$ waitress$^{0.10}$ got$^{0.09}$ took$^{0.08}$ order$^{0.08}$ asked$^{0.08}$ came$^{0.08}$ brought$^{0.07}$ finally$^{0.07}$ server$^{0.07}$
serviço+	great$^{0.24}$ place$^{0.14}$ restaurant$^{0.11}$ best$^{0.11}$ family$^{0.09}$ staff$^{0.09}$ friendly$^{0.09}$ excellent$^{0.08}$ atmosphere$^{0.08}$ food$^{0.08}$
valor-	just$^{0.17}$ waiter$^{0.09}$ served$^{0.06}$ wasn't$^{0.06}$ really$^{0.05}$ fries$^{0.05}$ restaurant$^{0.05}$ seafood$^{0.04}$ money$^{0.04}$ check$^{0.04}$
valor+	just$^{0.20}$ fresh$^{0.09}$ best$^{0.08}$ home$^{0.07}$ cream$^{0.05}$ excellent$^{0.05}$ waiter$^{0.05}$ pizza$^{0.05}$ hot$^{0.05}$ highly$^{0.05}$
ambiente-	food$^{0.69}$ eat$^{0.19}$ order$^{0.08}$ home$^{0.07}$ don't$^{0.05}$ owner$^{0.05}$ table$^{0.04}$ customers$^{0.04}$ pizza$^{0.04}$ chinese$^{0.03}$
ambiente+	food$^{0.69}$ eat$^{0.12}$ atmosphere$^{0.11}$ service$^{0.10}$ great$^{0.07}$ menu$^{0.06}$ home$^{0.06}$ ordered$^{0.05}$ dessert$^{0.05}$ burger$^{0.05}$
experiência-	ordered$^{0.13}$ said$^{0.12}$ pizza$^{0.10}$ waitress$^{0.09}$ took$^{0.08}$ manager$^{0.07}$ half$^{0.07}$ drinks$^{0.07}$ did$^{0.06}$ asked$^{0.06}$
experiência+	great$^{0.12}$ friendly$^{0.09}$ wonderful$^{0.08}$ seafood$^{0.08}$ excellent$^{0.08}$ perfect$^{0.07}$ best$^{0.07}$ delicious$^{0.06}$ quality$^{0.06}$ chef$^{0.06}$
10	like$^{0.46}$ good$^{0.29}$ experience$^{0.14}$ bar$^{0.13}$ price$^{0.11}$ small$^{0.11}$ meat$^{0.08}$ wait$^{0.08}$ lot$^{0.07}$ atmosphere$^{0.07}$
11	dinner$^{0.21}$ got$^{0.20}$ good$^{0.17}$ area$^{0.14}$ pretty$^{0.13}$ taste$^{0.09}$ didn't$^{0.09}$ great$^{0.09}$ experience$^{0.08}$ came$^{0.08}$
12	chicken$^{0.28}$ menu$^{0.23}$ place$^{0.21}$ salad$^{0.11}$ cheese$^{0.11}$ right$^{0.10}$ quite$^{0.10}$ served$^{0.08}$ i've$^{0.06}$ open$^{0.06}$
13	restaurant$^{0.52}$ time$^{0.27}$ meal$^{0.24}$ food$^{0.13}$ lunch$^{0.09}$ ordered$^{0.09}$ day$^{0.08}$ times$^{0.07}$ fish$^{0.07}$ bread$^{0.06}$
14	service$^{0.47}$ place$^{0.13}$ people$^{0.10}$ way$^{0.10}$ night$^{0.10}$ just$^{0.08}$ say$^{0.08}$ drink$^{0.07}$ looking$^{0.07}$ went$^{0.07}$

Tabela 22: Tópicos estimados com o modelo **Nota-Aspecto** e **20 tópicos** (5 avaliativos e10 não avaliativos) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	bad$^{0.15}$ terrible$^{0.09}$ bland$^{0.08}$ cold$^{0.08}$ told$^{0.08}$ ok$^{0.08}$ like$^{0.07}$ didn't$^{0.07}$ worst$^{0.07}$ money$^{0.06}$
comida+	delicious$^{0.14}$ good$^{0.11}$ nice$^{0.09}$ little$^{0.07}$ spicy$^{0.06}$ usually$^{0.06}$ make$^{0.06}$ variety$^{0.06}$ grilled$^{0.06}$ dishes$^{0.06}$
serviço-	minutes$^{0.23}$ did$^{0.16}$ waitress$^{0.15}$ took$^{0.11}$ got$^{0.10}$ server$^{0.10}$ finally$^{0.10}$ asked$^{0.10}$ wanted$^{0.09}$ slow$^{0.07}$
serviço+	great$^{0.36}$ excellent$^{0.13}$ best$^{0.12}$ friendly$^{0.12}$ staff$^{0.11}$ family$^{0.10}$ place$^{0.09}$ atmosphere$^{0.09}$ sauce$^{0.07}$ outstanding$^{0.06}$
valor-	did$^{0.17}$ table$^{0.10}$ waitress$^{0.10}$ wasn't$^{0.07}$ manager$^{0.07}$ waiter$^{0.06}$ wrong$^{0.05}$ burger$^{0.05}$ came$^{0.05}$ people$^{0.05}$
valor+	best$^{0.13}$ excellent$^{0.12}$ love$^{0.11}$ fresh$^{0.10}$ home$^{0.09}$ highly$^{0.08}$ favorite$^{0.08}$ wonderful$^{0.07}$ reasonable$^{0.07}$ homemade$^{0.06}$
ambiente-	owner$^{0.09}$ don't$^{0.09}$ order$^{0.07}$ eat$^{0.07}$ better$^{0.07}$ buffet$^{0.07}$ rude$^{0.06}$ table$^{0.06}$ told$^{0.06}$ customers$^{0.05}$
ambiente+	great$^{0.15}$ atmosphere$^{0.15}$ service$^{0.11}$ menu$^{0.11}$ delicious$^{0.09}$ dining$^{0.08}$ dessert$^{0.08}$ beautiful$^{0.06}$ burger$^{0.06}$ love$^{0.06}$
experiência-	said$^{0.22}$ came$^{0.15}$ ordered$^{0.14}$ didn't$^{0.14}$ left$^{0.12}$ asked$^{0.12}$ drinks$^{0.09}$ manager$^{0.09}$ told$^{0.07}$ horrible$^{0.07}$
experiência+	great$^{0.21}$ best$^{0.18}$ excellent$^{0.15}$ try$^{0.12}$ perfect$^{0.12}$ friendly$^{0.12}$ wonderful$^{0.10}$ fantastic$^{0.09}$ awesome$^{0.08}$ list$^{0.07}$
10	food$^{0.71}$ experience$^{0.31}$ it's$^{0.19}$ nice$^{0.14}$ home$^{0.12}$ sure$^{0.11}$ better$^{0.10}$ make$^{0.10}$ old$^{0.09}$ thing$^{0.06}$
11	food$^{0.33}$ dining$^{0.27}$ came$^{0.21}$ wait$^{0.20}$ table$^{0.12}$ order$^{0.10}$ times$^{0.08}$ appetizer$^{0.08}$ couple$^{0.07}$ fried$^{0.07}$
12	chicken$^{0.36}$ bar$^{0.26}$ meal$^{0.23}$ price$^{0.16}$ right$^{0.14}$ dinner$^{0.13}$ salad$^{0.13}$ soup$^{0.11}$ little$^{0.09}$ town$^{0.09}$
13	good$^{0.70}$ people$^{0.15}$ best$^{0.12}$ food$^{0.11}$ there's$^{0.10}$ meat$^{0.09}$ quality$^{0.08}$ return$^{0.07}$ friendly$^{0.07}$ eaten$^{0.07}$
14	like$^{0.54}$ restaurant$^{0.33}$ time$^{0.24}$ fries$^{0.18}$ atmosphere$^{0.14}$ sandwich$^{0.11}$ drinks$^{0.09}$ flavor$^{0.08}$ seafood$^{0.08}$ business$^{0.07}$
15	menu$^{0.30}$ small$^{0.21}$ just$^{0.20}$ got$^{0.19}$ time$^{0.17}$ good$^{0.17}$ ordered$^{0.13}$ inside$^{0.11}$ said$^{0.10}$ try$^{0.09}$
16	place$^{0.39}$ didn't$^{0.15}$ great$^{0.14}$ special$^{0.13}$ pretty$^{0.13}$ area$^{0.12}$ night$^{0.10}$ steak$^{0.10}$ meal$^{0.08}$ dinner$^{0.08}$
17	service$^{0.38}$ food$^{0.27}$ eat$^{0.27}$ cheese$^{0.17}$ way$^{0.16}$ went$^{0.16}$ quite$^{0.16}$ looking$^{0.10}$ bit$^{0.09}$ want$^{0.08}$
18	restaurant$^{0.42}$ lunch$^{0.13}$ i'm$^{0.11}$ little$^{0.09}$ room$^{0.09}$ menu$^{0.09}$ ordered$^{0.08}$ know$^{0.08}$ open$^{0.08}$ fine$^{0.08}$
19	really$^{0.44}$ service$^{0.29}$ pizza$^{0.22}$ sauce$^{0.18}$ say$^{0.12}$ going$^{0.11}$ place$^{0.10}$ wine$^{0.08}$ don't$^{0.08}$ places$^{0.08}$

Tabela 23: Tópicos estimados com o modelo **Nota-Aspecto** e **30 tópicos** (5 avaliativos e 20 não avaliativos) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	bad$^{0.20}$ cold$^{0.14}$ terrible$^{0.13}$ worst$^{0.12}$ bland$^{0.12}$ told$^{0.12}$ ok$^{0.11}$ dry$^{0.09}$ tasted$^{0.07}$ overcooked$^{0.07}$
comida+	wonderful$^{0.17}$ great$^{0.13}$ fresh$^{0.11}$ dishes$^{0.11}$ incredible$^{0.09}$ favorite$^{0.09}$ spicy$^{0.09}$ loved$^{0.09}$ grilled$^{0.08}$ italian$^{0.08}$
serviço-	minutes$^{0.37}$ took$^{0.28}$ waitress$^{0.26}$ finally$^{0.16}$ waited$^{0.14}$ asked$^{0.12}$ brought$^{0.11}$ tip$^{0.10}$ wanted$^{0.10}$ received$^{0.09}$
serviço+	great$^{0.58}$ friendly$^{0.21}$ staff$^{0.19}$ excellent$^{0.19}$ best$^{0.14}$ world$^{0.13}$ amazing$^{0.11}$ outstanding$^{0.09}$ family$^{0.08}$ fabulous$^{0.08}$
valor-	did$^{0.39}$ waitress$^{0.12}$ manager$^{0.11}$ money$^{0.10}$ arrived$^{0.09}$ rude$^{0.08}$ wrong$^{0.08}$ ordered$^{0.07}$ paid$^{0.06}$ seated$^{0.06}$
valor+	love$^{0.23}$ best$^{0.21}$ excellent$^{0.19}$ try$^{0.18}$ fresh$^{0.13}$ reasonable$^{0.11}$ home$^{0.10}$ favorite$^{0.10}$ french$^{0.10}$ homemade$^{0.08}$
ambiente-	better$^{0.14}$ order$^{0.13}$ rude$^{0.13}$ owner$^{0.11}$ told$^{0.10}$ customers$^{0.08}$ buffet$^{0.06}$ don't$^{0.05}$ tables$^{0.05}$ dirty$^{0.05}$
ambiente+	great$^{0.31}$ atmosphere$^{0.25}$ wonderful$^{0.18}$ menu$^{0.13}$ delicious$^{0.12}$ dessert$^{0.11}$ nice$^{0.11}$ beautiful$^{0.10}$ unique$^{0.10}$ topped$^{0.08}$
experiência-	said$^{0.30}$ didn't$^{0.29}$ asked$^{0.24}$ came$^{0.15}$ left$^{0.13}$ manager$^{0.13}$ poor$^{0.11}$ later$^{0.10}$ waitress$^{0.09}$ bad$^{0.09}$
experiência+	excellent$^{0.24}$ highly$^{0.19}$ great$^{0.19}$ perfect$^{0.18}$ delicious$^{0.14}$ best$^{0.13}$ fantastic$^{0.11}$ enjoy$^{0.09}$ sweet$^{0.08}$ friendly$^{0.08}$
10	price$^{0.21}$ good$^{0.16}$ got$^{0.14}$ appetizer$^{0.13}$ sauce$^{0.10}$ used$^{0.10}$ rice$^{0.09}$ evening$^{0.09}$ know$^{0.08}$ seated$^{0.08}$
11	just$^{0.38}$ sandwich$^{0.17}$ come$^{0.17}$ hot$^{0.16}$ night$^{0.14}$ steak$^{0.13}$ waiter$^{0.13}$ table$^{0.10}$ people$^{0.09}$ came$^{0.08}$
12	meal$^{0.41}$ experience$^{0.25}$ food$^{0.21}$ eat$^{0.10}$ good$^{0.10}$ make$^{0.09}$ got$^{0.09}$ bread$^{0.08}$ house$^{0.07}$ staff$^{0.07}$
13	food$^{0.84}$ fries$^{0.17}$ soup$^{0.17}$ times$^{0.15}$ ordered$^{0.13}$ town$^{0.10}$ cooked$^{0.10}$ husband$^{0.06}$ grill$^{0.06}$ friends$^{0.06}$
14	experience$^{0.28}$ served$^{0.20}$ atmosphere$^{0.18}$ dining$^{0.15}$ service$^{0.14}$ right$^{0.12}$ order$^{0.09}$ long$^{0.09}$ fine$^{0.09}$ friendly$^{0.08}$
15	really$^{0.39}$ people$^{0.29}$ good$^{0.19}$ table$^{0.12}$ open$^{0.12}$ like$^{0.12}$ little$^{0.11}$ nice$^{0.10}$ quality$^{0.10}$ fish$^{0.09}$
16	good$^{0.54}$ cheese$^{0.14}$ time$^{0.13}$ wife$^{0.10}$ couple$^{0.09}$ sure$^{0.09}$ seating$^{0.08}$ did$^{0.08}$ best$^{0.08}$ dinner$^{0.07}$
17	wait$^{0.24}$ cream$^{0.13}$ seafood$^{0.13}$ like$^{0.11}$ got$^{0.11}$ right$^{0.10}$ visit$^{0.09}$ great$^{0.09}$ served$^{0.09}$ chips$^{0.09}$
18	it's$^{0.52}$ bar$^{0.35}$ dinner$^{0.12}$ fries$^{0.12}$ didn't$^{0.11}$ sure$^{0.09}$ time$^{0.09}$ night$^{0.08}$ offer$^{0.08}$ local$^{0.08}$
19	dining$^{0.45}$ don't$^{0.22}$ chicken$^{0.19}$ make$^{0.18}$ check$^{0.16}$ busy$^{0.12}$ day$^{0.12}$ area$^{0.10}$ return$^{0.08}$ atmosphere$^{0.08}$
20	area$^{0.20}$ food$^{0.18}$ cheese$^{0.18}$ good$^{0.11}$ meal$^{0.11}$ waitress$^{0.11}$ years$^{0.10}$ can't$^{0.09}$ chicken$^{0.08}$ eaten$^{0.08}$
21	like$^{0.37}$ better$^{0.29}$ time$^{0.29}$ came$^{0.24}$ place$^{0.21}$ salad$^{0.19}$ eat$^{0.13}$ lot$^{0.09}$ just$^{0.07}$ kitchen$^{0.06}$
22	menu$^{0.52}$ home$^{0.25}$ special$^{0.23}$ say$^{0.19}$ said$^{0.13}$ steak$^{0.12}$ just$^{0.11}$ salad$^{0.11}$ 2$^{0.08}$ half$^{0.07}$
23	restaurant$^{0.50}$ small$^{0.26}$ place$^{0.24}$ great$^{0.15}$ really$^{0.08}$ location$^{0.08}$ tables$^{0.08}$ plate$^{0.08}$ way$^{0.07}$ going$^{0.07}$
24	nice$^{0.16}$ chicken$^{0.15}$ know$^{0.14}$ ordered$^{0.13}$ new$^{0.11}$ tried$^{0.10}$ actually$^{0.10}$ year$^{0.09}$ decor$^{0.08}$ burger$^{0.07}$
25	food$^{0.79}$ time$^{0.31}$ like$^{0.16}$ way$^{0.15}$ great$^{0.14}$ i'm$^{0.11}$ hard$^{0.10}$ dinner$^{0.10}$ did$^{0.10}$ looking$^{0.09}$
26	food$^{0.42}$ sauce$^{0.31}$ pretty$^{0.26}$ quite$^{0.20}$ shrimp$^{0.16}$ i've$^{0.14}$ prices$^{0.12}$ thought$^{0.11}$ worth$^{0.10}$ called$^{0.07}$
27	restaurant$^{1.05}$ service$^{0.45}$ restaurants$^{0.12}$ bit$^{0.12}$ meals$^{0.09}$ dish$^{0.08}$ went$^{0.08}$ wine$^{0.06}$ come$^{0.06}$ potato$^{0.05}$
28	service$^{0.52}$ lunch$^{0.28}$ room$^{0.16}$ don't$^{0.13}$ dessert$^{0.12}$ red$^{0.12}$ staff$^{0.11}$ large$^{0.10}$ menu$^{0.09}$ restaurant$^{0.08}$
29	place$^{0.85}$ pizza$^{0.42}$ drink$^{0.19}$ little$^{0.17}$ wine$^{0.12}$ table$^{0.11}$ crab$^{0.10}$ italian$^{0.08}$ dishes$^{0.07}$ salad$^{0.07}$

Tabela 24: Tópicos estimados com o modelo **Nota-Aspecto** e **50 tópicos** (5 avaliativos e 40 não avaliativos) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	bad$^{0.22}$ terrible$^{0.21}$ didn't$^{0.19}$ bland$^{0.18}$ ok$^{0.18}$ cold$^{0.13}$ told$^{0.13}$ dry$^{0.12}$ overcooked$^{0.11}$ worst$^{0.09}$
comida+	delicious$^{0.39}$ wonderful$^{0.28}$ great$^{0.24}$ chocolate$^{0.17}$ favorite$^{0.16}$ love$^{0.13}$ fresh$^{0.13}$ spicy$^{0.11}$ loved$^{0.10}$ beef$^{0.09}$
serviço-	minutes$^{0.57}$ waitress$^{0.29}$ took$^{0.24}$ finally$^{0.22}$ left$^{0.22}$ waiter$^{0.21}$ waited$^{0.18}$ hour$^{0.10}$ worst$^{0.10}$ 20$^{0.09}$
serviço+	great$^{0.99}$ friendly$^{0.42}$ best$^{0.33}$ amazing$^{0.17}$ excellent$^{0.16}$ world$^{0.13}$ fun$^{0.10}$ fabulous$^{0.10}$ awesome$^{0.08}$ italian$^{0.07}$
valor-	did$^{0.50}$ manager$^{0.22}$ minutes$^{0.15}$ charged$^{0.14}$ money$^{0.12}$ pay$^{0.10}$ seated$^{0.09}$ rude$^{0.09}$ overpriced$^{0.09}$ wasn't$^{0.08}$
valor+	best$^{0.60}$ fresh$^{0.27}$ excellent$^{0.26}$ love$^{0.22}$ great$^{0.19}$ favorite$^{0.17}$ reasonable$^{0.15}$ recommend$^{0.13}$ homemade$^{0.11}$ generous$^{0.08}$
ambiente-	rude$^{0.21}$ owner$^{0.16}$ told$^{0.15}$ customers$^{0.14}$ arrived$^{0.12}$ asked$^{0.10}$ wrong$^{0.08}$ dirty$^{0.08}$ better$^{0.08}$ charge$^{0.07}$
ambiente+	great$^{0.72}$ atmosphere$^{0.33}$ wonderful$^{0.32}$ dessert$^{0.20}$ beautiful$^{0.18}$ excellent$^{0.13}$ awesome$^{0.11}$ pub$^{0.08}$ love$^{0.07}$ unique$^{0.07}$
experiência-	said$^{0.35}$ asked$^{0.27}$ bad$^{0.21}$ poor$^{0.21}$ horrible$^{0.20}$ told$^{0.15}$ came$^{0.15}$ manager$^{0.15}$ awful$^{0.13}$ waitress$^{0.12}$
experiência+	excellent$^{0.38}$ highly$^{0.29}$ perfect$^{0.28}$ fantastic$^{0.23}$ outstanding$^{0.18}$ staff$^{0.15}$ incredible$^{0.13}$ cozy$^{0.11}$ delicious$^{0.09}$ list$^{0.09}$
10	sauce$^{0.21}$ it's$^{0.19}$ know$^{0.14}$ didn't$^{0.13}$ dining$^{0.12}$ left$^{0.11}$ pleasant$^{0.09}$ end$^{0.08}$ just$^{0.08}$ plate$^{0.07}$
11	menu$^{0.28}$ try$^{0.17}$ table$^{0.17}$ it's$^{0.13}$ best$^{0.11}$ make$^{0.10}$ style$^{0.09}$ fries$^{0.08}$ offered$^{0.08}$ place$^{0.07}$
12	nice$^{0.30}$ best$^{0.29}$ just$^{0.28}$ ordered$^{0.28}$ came$^{0.21}$ soup$^{0.21}$ wanted$^{0.07}$ sandwich$^{0.07}$ entrees$^{0.07}$ delicious$^{0.07}$
13	good$^{0.26}$ people$^{0.19}$ experience$^{0.13}$ seated$^{0.10}$ second$^{0.10}$ big$^{0.09}$ pizza$^{0.08}$ rice$^{0.08}$ going$^{0.08}$ husband$^{0.07}$
14	good$^{1.23}$ bread$^{0.15}$ items$^{0.09}$ experience$^{0.08}$ main$^{0.08}$ used$^{0.07}$ italian$^{0.06}$ hotel$^{0.06}$ tasted$^{0.06}$ soup$^{0.05}$
15	people$^{0.27}$ think$^{0.26}$ make$^{0.20}$ staff$^{0.17}$ time$^{0.12}$ tables$^{0.12}$ cooked$^{0.11}$ restaurants$^{0.10}$ family$^{0.09}$ owner$^{0.09}$
16	food$^{0.98}$ got$^{0.32}$ just$^{0.31}$ table$^{0.13}$ pay$^{0.08}$ serve$^{0.08}$ dining$^{0.07}$ did$^{0.07}$ quick$^{0.07}$ experience$^{0.06}$
17	new$^{0.38}$ service$^{0.35}$ evening$^{0.16}$ food$^{0.15}$ wife$^{0.12}$ ate$^{0.10}$ large$^{0.09}$ eating$^{0.08}$ salad$^{0.07}$ roll$^{0.07}$
18	night$^{0.26}$ served$^{0.18}$ kind$^{0.18}$ thing$^{0.17}$ try$^{0.14}$ visit$^{0.11}$ didn't$^{0.10}$ staff$^{0.09}$ cheese$^{0.07}$ fresh$^{0.07}$
19	food$^{0.46}$ times$^{0.25}$ price$^{0.22}$ french$^{0.11}$ chicken$^{0.09}$ dishes$^{0.08}$ eat$^{0.07}$ inside$^{0.07}$ hot$^{0.07}$ set$^{0.06}$
20	food$^{0.65}$ cheese$^{0.26}$ wait$^{0.24}$ want$^{0.21}$ quality$^{0.10}$ did$^{0.09}$ looked$^{0.09}$ look$^{0.07}$ lot$^{0.07}$ don't$^{0.06}$
21	meal$^{0.57}$ really$^{0.45}$ served$^{0.15}$ got$^{0.14}$ hot$^{0.14}$ return$^{0.11}$ said$^{0.10}$ years$^{0.09}$ table$^{0.09}$ restaurant$^{0.09}$
22	come$^{0.33}$ food$^{0.32}$ fries$^{0.18}$ eaten$^{0.13}$ area$^{0.11}$ spot$^{0.10}$ bar$^{0.10}$ reasonable$^{0.10}$ chips$^{0.09}$ served$^{0.09}$
23	order$^{0.32}$ experience$^{0.22}$ said$^{0.12}$ waiter$^{0.11}$ long$^{0.10}$ cooking$^{0.09}$ friendly$^{0.08}$ ordered$^{0.08}$ 20$^{0.07}$ eat$^{0.06}$
24	menu$^{0.34}$ cheese$^{0.32}$ open$^{0.22}$ fries$^{0.14}$ restaurant$^{0.11}$ chef$^{0.10}$ recommend$^{0.09}$ start$^{0.07}$ friendly$^{0.07}$ sure$^{0.07}$
25	really$^{0.21}$ restaurant$^{0.17}$ wait$^{0.16}$ got$^{0.15}$ potatoes$^{0.09}$ buffet$^{0.09}$ little$^{0.09}$ make$^{0.08}$ selection$^{0.08}$ don't$^{0.07}$
26	little$^{0.29}$ restaurant$^{0.22}$ eat$^{0.15}$ right$^{0.15}$ chicken$^{0.14}$ it's$^{0.10}$ came$^{0.10}$ people$^{0.08}$ dining$^{0.08}$ way$^{0.07}$
27	bar$^{0.31}$ like$^{0.29}$ it's$^{0.20}$ better$^{0.18}$ visit$^{0.13}$ server$^{0.12}$ friends$^{0.12}$ time$^{0.10}$ brought$^{0.08}$ return$^{0.08}$
28	dinner$^{0.23}$ special$^{0.21}$ menu$^{0.20}$ atmosphere$^{0.18}$ chicken$^{0.17}$ years$^{0.16}$ bar$^{0.14}$ night$^{0.14}$ try$^{0.13}$ family$^{0.08}$
29	experience$^{0.28}$ good$^{0.25}$ make$^{0.16}$ like$^{0.16}$ old$^{0.15}$ cream$^{0.11}$ fish$^{0.11}$ fresh$^{0.11}$ it's$^{0.10}$ small$^{0.07}$
30	huge$^{0.20}$ time$^{0.19}$ sandwich$^{0.16}$ menu$^{0.14}$ room$^{0.12}$ town$^{0.12}$ brought$^{0.12}$ better$^{0.09}$ area$^{0.08}$ red$^{0.08}$
31	place$^{0.35}$ small$^{0.30}$ good$^{0.18}$ restaurant$^{0.14}$ time$^{0.12}$ ordered$^{0.12}$ right$^{0.11}$ local$^{0.09}$ little$^{0.08}$ baked$^{0.08}$
32	salad$^{0.31}$ day$^{0.22}$ just$^{0.22}$ area$^{0.20}$ couple$^{0.18}$ like$^{0.13}$ offer$^{0.13}$ half$^{0.10}$ water$^{0.10}$ bread$^{0.09}$
33	came$^{0.25}$ dining$^{0.25}$ good$^{0.24}$ going$^{0.18}$ sandwich$^{0.12}$ time$^{0.12}$ chain$^{0.09}$ check$^{0.07}$ meal$^{0.07}$ away$^{0.06}$
34	place$^{0.33}$ really$^{0.23}$ burger$^{0.16}$ chicken$^{0.13}$ got$^{0.13}$ lunch$^{0.10}$ know$^{0.10}$ big$^{0.10}$ sushi$^{0.07}$ heard$^{0.07}$
35	ordered$^{0.26}$ steak$^{0.19}$ restaurant$^{0.17}$ don't$^{0.15}$ old$^{0.11}$ pasta$^{0.09}$ flavor$^{0.09}$ sandwiches$^{0.07}$ way$^{0.07}$ pretty$^{0.07}$
36	food$^{0.36}$ seafood$^{0.31}$ plate$^{0.15}$ dish$^{0.15}$ place$^{0.12}$ party$^{0.11}$ home$^{0.10}$ 2$^{0.10}$ salad$^{0.09}$ nice$^{0.08}$
37	like$^{0.37}$ nice$^{0.26}$ restaurant$^{0.24}$ went$^{0.22}$ took$^{0.20}$ taste$^{0.12}$ sunday$^{0.08}$ inside$^{0.07}$ large$^{0.07}$ shrimp$^{0.06}$
38	cafe$^{0.24}$ wine$^{0.23}$ home$^{0.21}$ didn't$^{0.21}$ did$^{0.18}$ 3$^{0.10}$ waitress$^{0.10}$ experience$^{0.10}$ places$^{0.08}$ right$^{0.07}$
39	service$^{0.81}$ food$^{0.35}$ business$^{0.12}$ order$^{0.09}$ looking$^{0.09}$ potato$^{0.08}$ meals$^{0.06}$ husband$^{0.06}$ portions$^{0.06}$ try$^{0.05}$
40	like$^{0.34}$ meal$^{0.27}$ sauce$^{0.21}$ really$^{0.13}$ order$^{0.11}$ decided$^{0.10}$ say$^{0.09}$ large$^{0.08}$ hard$^{0.07}$ half$^{0.07}$
41	time$^{0.56}$ just$^{0.44}$ eat$^{0.21}$ quite$^{0.14}$ good$^{0.09}$ sandwich$^{0.09}$ saturday$^{0.08}$ need$^{0.07}$ sweet$^{0.07}$ makes$^{0.06}$
42	good$^{0.51}$ great$^{0.43}$ say$^{0.22}$ sauce$^{0.14}$ eat$^{0.13}$ salad$^{0.10}$ place$^{0.10}$ ask$^{0.09}$ sushi$^{0.07}$ little$^{0.07}$
43	pretty$^{0.33}$ great$^{0.31}$ service$^{0.27}$ way$^{0.13}$ table$^{0.08}$ thought$^{0.08}$ green$^{0.08}$ bit$^{0.06}$ reason$^{0.06}$ changed$^{0.05}$
44	food$^{0.42}$ don't$^{0.20}$ tried$^{0.15}$ quite$^{0.14}$ i'm$^{0.14}$ breakfast$^{0.13}$ it's$^{0.11}$ near$^{0.11}$ new$^{0.11}$ went$^{0.10}$
45	service$^{0.75}$ dessert$^{0.17}$ 2$^{0.11}$ times$^{0.08}$ use$^{0.08}$ lunch$^{0.08}$ actually$^{0.08}$ glass$^{0.07}$ having$^{0.07}$ small$^{0.06}$
46	restaurant$^{0.50}$ place$^{0.44}$ i've$^{0.41}$ like$^{0.09}$ care$^{0.09}$ drinks$^{0.09}$ think$^{0.09}$ chef$^{0.08}$ pretty$^{0.06}$ bread$^{0.06}$
47	pizza$^{0.26}$ staff$^{0.19}$ served$^{0.17}$ restaurants$^{0.16}$ meals$^{0.13}$ worth$^{0.13}$ order$^{0.12}$ buffet$^{0.12}$ just$^{0.11}$ come$^{0.11}$
48	place$^{0.47}$ lunch$^{0.41}$ drink$^{0.27}$ dinner$^{0.22}$ room$^{0.21}$ prices$^{0.16}$ menu$^{0.13}$ ordered$^{0.07}$ lot$^{0.07}$ went$^{0.07}$
49	pizza$^{0.39}$ wine$^{0.22}$ table$^{0.15}$ food$^{0.15}$ located$^{0.14}$ italian$^{0.13}$ minutes$^{0.09}$ salad$^{0.09}$ crab$^{0.08}$ told$^{0.07}$

Tabela 25: Tópicos estimados com o modelo **Nota-Aspecto LDA** e **10 tópicos** (5 avaliativos e nenhum não avaliativo) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
0	food$^{0.54}$ good$^{0.18}$ service$^{0.14}$ place$^{0.12}$ great$^{0.11}$ atmosphere$^{0.06}$ staff$^{0.05}$ like$^{0.05}$ best$^{0.05}$ better$^{0.04}$
1	food$^{0.60}$ great$^{0.46}$ good$^{0.25}$ service$^{0.21}$ place$^{0.19}$ excellent$^{0.13}$ atmosphere$^{0.12}$ friendly$^{0.08}$ nice$^{0.07}$ staff$^{0.06}$
2	minutes$^{0.11}$ table$^{0.10}$ waitress$^{0.08}$ waiter$^{0.08}$ service$^{0.07}$ order$^{0.06}$ took$^{0.06}$ server$^{0.06}$ experience$^{0.06}$ seated$^{0.05}$
3	experience$^{0.13}$ restaurant$^{0.06}$ dinner$^{0.06}$ service$^{0.05}$ wine$^{0.05}$ wonderful$^{0.05}$ staff$^{0.05}$ friendly$^{0.05}$ food$^{0.05}$ table$^{0.04}$
4	really$^{0.11}$ got$^{0.08}$ good$^{0.08}$ little$^{0.07}$ just$^{0.06}$ pretty$^{0.06}$ fries$^{0.06}$ pizza$^{0.06}$ came$^{0.06}$ it's$^{0.06}$
5	little$^{0.09}$ really$^{0.09}$ just$^{0.07}$ good$^{0.07}$ pizza$^{0.06}$ got$^{0.06}$ fries$^{0.05}$ sandwich$^{0.05}$ came$^{0.04}$ pretty$^{0.04}$
6	chicken$^{0.05}$ served$^{0.05}$ salad$^{0.04}$ buffet$^{0.04}$ sauce$^{0.04}$ shrimp$^{0.04}$ fried$^{0.03}$ dish$^{0.03}$ rice$^{0.03}$ seafood$^{0.03}$
7	chicken$^{0.06}$ fresh$^{0.06}$ served$^{0.05}$ salad$^{0.05}$ shrimp$^{0.05}$ seafood$^{0.04}$ menu$^{0.04}$ sauce$^{0.04}$ dish$^{0.03}$ cream$^{0.03}$
8	eat$^{0.08}$ said$^{0.08}$ just$^{0.06}$ told$^{0.06}$ restaurant$^{0.05}$ know$^{0.05}$ didn't$^{0.05}$ don't$^{0.04}$ people$^{0.04}$ way$^{0.04}$
9	best$^{0.12}$ new$^{0.06}$ years$^{0.06}$ restaurant$^{0.06}$ just$^{0.06}$ eat$^{0.06}$ sushi$^{0.05}$ try$^{0.05}$ chef$^{0.05}$ like$^{0.04}$

Tabela 26: Tópicos estimados com o modelo **Nota-Aspecto LDA** e **15 tópicos** (5 avaliativos e 5 não avaliativos) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	chicken$^{0.14}$ salad$^{0.11}$ ordered$^{0.10}$ served$^{0.10}$ sauce$^{0.09}$ shrimp$^{0.09}$ bread$^{0.08}$ soup$^{0.07}$ tasted$^{0.07}$ fried$^{0.06}$
comida+	chicken$^{0.15}$ served$^{0.12}$ sauce$^{0.09}$ salad$^{0.08}$ soup$^{0.07}$ cream$^{0.06}$ bread$^{0.06}$ fried$^{0.05}$ chocolate$^{0.05}$ garlic$^{0.05}$
serviço-	minutes$^{0.23}$ food$^{0.15}$ took$^{0.15}$ service$^{0.15}$ table$^{0.15}$ waitress$^{0.13}$ order$^{0.13}$ waiter$^{0.09}$ got$^{0.09}$ did$^{0.09}$
serviço+	great$^{0.83}$ food$^{0.34}$ service$^{0.24}$ place$^{0.21}$ atmosphere$^{0.20}$ friendly$^{0.20}$ excellent$^{0.18}$ staff$^{0.13}$ wonderful$^{0.11}$ recommend$^{0.08}$
valor-	menu$^{0.20}$ buffet$^{0.12}$ seafood$^{0.11}$ fish$^{0.09}$ lunch$^{0.08}$ items$^{0.07}$ large$^{0.06}$ sushi$^{0.05}$ dishes$^{0.05}$ fresh$^{0.05}$
valor+	menu$^{0.16}$ fresh$^{0.12}$ seafood$^{0.12}$ sushi$^{0.12}$ dishes$^{0.11}$ buffet$^{0.10}$ fish$^{0.08}$ breakfast$^{0.08}$ lunch$^{0.08}$ variety$^{0.07}$
ambiente-	restaurant$^{0.24}$ dining$^{0.16}$ wine$^{0.09}$ room$^{0.07}$ new$^{0.06}$ tables$^{0.06}$ bar$^{0.06}$ better$^{0.05}$ experience$^{0.05}$ kitchen$^{0.04}$
ambiente+	dining$^{0.23}$ restaurant$^{0.19}$ wine$^{0.17}$ experience$^{0.11}$ chef$^{0.07}$ list$^{0.07}$ room$^{0.06}$ fine$^{0.05}$ beautiful$^{0.05}$ bar$^{0.05}$
experiência-	said$^{0.17}$ manager$^{0.15}$ bad$^{0.12}$ told$^{0.11}$ owner$^{0.11}$ rude$^{0.09}$ did$^{0.09}$ asked$^{0.07}$ waitress$^{0.07}$ business$^{0.07}$
experiência+	best$^{0.41}$ restaurant$^{0.18}$ years$^{0.13}$ try$^{0.10}$ love$^{0.10}$ new$^{0.10}$ favorite$^{0.09}$ family$^{0.07}$ visit$^{0.07}$ i've$^{0.06}$
10	pizza$^{0.25}$ good$^{0.21}$ place$^{0.10}$ like$^{0.10}$ bbq$^{0.09}$ little$^{0.09}$ it's$^{0.09}$ i've$^{0.07}$ beef$^{0.07}$ really$^{0.07}$
11	dinner$^{0.16}$ time$^{0.12}$ meal$^{0.12}$ ordered$^{0.12}$ table$^{0.12}$ night$^{0.11}$ husband$^{0.10}$ wife$^{0.07}$ wait$^{0.07}$ waiter$^{0.06}$
12	food$^{0.97}$ good$^{0.47}$ service$^{0.29}$ great$^{0.25}$ restaurant$^{0.13}$ atmosphere$^{0.13}$ price$^{0.12}$ mexican$^{0.08}$ times$^{0.08}$ staff$^{0.07}$
13	place$^{0.27}$ just$^{0.22}$ time$^{0.18}$ don't$^{0.17}$ eat$^{0.17}$ like$^{0.14}$ know$^{0.13}$ it's$^{0.12}$ want$^{0.12}$ going$^{0.11}$
14	really$^{0.13}$ got$^{0.13}$ fries$^{0.10}$ little$^{0.09}$ came$^{0.09}$ sandwich$^{0.09}$ burger$^{0.08}$ didn't$^{0.08}$ pretty$^{0.06}$ lot$^{0.05}$

Tabela 27: Tópicos estimados com o modelo **Nota-Aspecto LDA** e **30 tópicos** (5 avaliativos e 20 não avaliativos) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	bad$^{0.31}$ ordered$^{0.20}$ terrible$^{0.15}$ worst$^{0.14}$ bland$^{0.13}$ horrible$^{0.13}$ tasted$^{0.12}$ sushi$^{0.09}$ cold$^{0.09}$ ok$^{0.09}$
comida+	sushi$^{0.34}$ fresh$^{0.27}$ chef$^{0.16}$ delicious$^{0.10}$ dishes$^{0.10}$ incredible$^{0.10}$ roll$^{0.09}$ menu$^{0.09}$ spicy$^{0.07}$ rolls$^{0.06}$
serviço-	minutes$^{0.37}$ waitress$^{0.23}$ order$^{0.21}$ took$^{0.20}$ waiter$^{0.15}$ waited$^{0.15}$ finally$^{0.14}$ table$^{0.14}$ got$^{0.14}$ brought$^{0.13}$
serviço+	best$^{0.52}$ excellent$^{0.26}$ family$^{0.25}$ friendly$^{0.21}$ staff$^{0.15}$ italian$^{0.15}$ prices$^{0.13}$ world$^{0.12}$ outstanding$^{0.12}$ home$^{0.11}$
valor-	didn't$^{0.12}$ better$^{0.10}$ given$^{0.09}$ tables$^{0.09}$ overpriced$^{0.08}$ money$^{0.08}$ think$^{0.07}$ ok$^{0.06}$ charge$^{0.06}$ patrons$^{0.05}$
valor+	dining$^{0.07}$ french$^{0.07}$ you'll$^{0.06}$ it's$^{0.06}$ walls$^{0.06}$ friday$^{0.05}$ selections$^{0.05}$ south$^{0.05}$ dessert$^{0.04}$ different$^{0.04}$
ambiente-	owner$^{0.26}$ rude$^{0.17}$ customers$^{0.15}$ card$^{0.09}$ credit$^{0.09}$ kitchen$^{0.09}$ business$^{0.07}$ don't$^{0.06}$ children$^{0.06}$ problem$^{0.05}$
ambiente+	atmosphere$^{0.29}$ wonderful$^{0.27}$ service$^{0.24}$ excellent$^{0.17}$ delicious$^{0.15}$ beautiful$^{0.13}$ nice$^{0.12}$ cozy$^{0.09}$ great$^{0.08}$ selection$^{0.08}$
experiência-	said$^{0.36}$ manager$^{0.31}$ asked$^{0.28}$ told$^{0.25}$ did$^{0.18}$ service$^{0.13}$ waitress$^{0.13}$ left$^{0.12}$ came$^{0.12}$ tip$^{0.10}$
experiência+	great$^{1.09}$ love$^{0.31}$ excellent$^{0.26}$ highly$^{0.24}$ recommend$^{0.24}$ best$^{0.23}$ wonderful$^{0.22}$ amazing$^{0.17}$ awesome$^{0.15}$ fantastic$^{0.15}$
10	pizza$^{0.64}$ good$^{0.18}$ really$^{0.17}$ little$^{0.17}$ pretty$^{0.16}$ cheese$^{0.13}$ chicago$^{0.12}$ order$^{0.11}$ just$^{0.09}$ got$^{0.08}$
11	bar$^{0.79}$ room$^{0.23}$ good$^{0.19}$ area$^{0.15}$ really$^{0.14}$ nice$^{0.13}$ check$^{0.11}$ beer$^{0.11}$ dining$^{0.10}$ night$^{0.09}$
12	experience$^{0.82}$ restaurant$^{0.67}$ food$^{0.38}$ staff$^{0.28}$ service$^{0.27}$ family$^{0.19}$ dining$^{0.19}$ recommend$^{0.16}$ wife$^{0.15}$ pleasant$^{0.11}$
13	new$^{0.43}$ years$^{0.33}$ restaurant$^{0.26}$ times$^{0.18}$ try$^{0.15}$ visit$^{0.13}$ year$^{0.13}$ review$^{0.11}$ location$^{0.10}$ old$^{0.10}$
14	just$^{0.35}$ know$^{0.33}$ don't$^{0.31}$ want$^{0.30}$ people$^{0.21}$ make$^{0.18}$ sure$^{0.17}$ like$^{0.17}$ way$^{0.13}$ going$^{0.13}$
15	chicken$^{0.54}$ dishes$^{0.25}$ fried$^{0.21}$ cream$^{0.21}$ rice$^{0.19}$ dish$^{0.14}$ ice$^{0.14}$ hot$^{0.14}$ chinese$^{0.11}$ soup$^{0.11}$
16	bbq$^{0.24}$ sauce$^{0.21}$ pork$^{0.19}$ meat$^{0.14}$ good$^{0.14}$ ribs$^{0.12}$ flavor$^{0.10}$ just$^{0.09}$ grill$^{0.09}$ home$^{0.09}$
17	it's$^{0.56}$ place$^{0.49}$ eat$^{0.39}$ i've$^{0.30}$ best$^{0.28}$ food$^{0.21}$ eaten$^{0.18}$ time$^{0.15}$ places$^{0.13}$ bad$^{0.12}$
18	got$^{0.20}$ fries$^{0.18}$ really$^{0.18}$ sandwich$^{0.17}$ burger$^{0.17}$ came$^{0.16}$ didn't$^{0.11}$ cheese$^{0.09}$ wasn't$^{0.07}$ pretty$^{0.06}$
19	buffet$^{0.30}$ breakfast$^{0.28}$ lunch$^{0.27}$ cafe$^{0.19}$ day$^{0.15}$ coffee$^{0.13}$ menu$^{0.12}$ local$^{0.11}$ diner$^{0.11}$ large$^{0.10}$
20	little$^{0.32}$ just$^{0.13}$ inside$^{0.11}$ door$^{0.11}$ building$^{0.11}$ right$^{0.11}$ like$^{0.09}$ look$^{0.09}$ outside$^{0.09}$ street$^{0.09}$
21	good$^{1.36}$ food$^{0.75}$ service$^{0.47}$ price$^{0.33}$ prices$^{0.23}$ quality$^{0.21}$ worth$^{0.13}$ portions$^{0.12}$ excellent$^{0.11}$ reasonable$^{0.11}$
22	steak$^{0.35}$ seafood$^{0.34}$ shrimp$^{0.34}$ fish$^{0.24}$ crab$^{0.24}$ fresh$^{0.18}$ house$^{0.09}$ salmon$^{0.08}$ lobster$^{0.08}$ menu$^{0.08}$
23	wine$^{0.47}$ menu$^{0.45}$ special$^{0.27}$ restaurant$^{0.25}$ dinner$^{0.23}$ list$^{0.17}$ evening$^{0.12}$ dining$^{0.08}$ specials$^{0.07}$ fine$^{0.07}$
24	wait$^{0.53}$ time$^{0.26}$ table$^{0.21}$ people$^{0.19}$ 2$^{0.18}$ party$^{0.14}$ busy$^{0.14}$ seated$^{0.13}$ minutes$^{0.12}$ reservations$^{0.11}$
25	ordered$^{0.32}$ went$^{0.25}$ time$^{0.24}$ husband$^{0.23}$ dinner$^{0.19}$ did$^{0.15}$ going$^{0.12}$ meal$^{0.12}$ thought$^{0.12}$ didn't$^{0.12}$
26	dessert$^{0.17}$ table$^{0.15}$ course$^{0.12}$ glass$^{0.12}$ water$^{0.12}$ waiter$^{0.11}$ quite$^{0.10}$ restaurant$^{0.09}$ server$^{0.08}$ appetizer$^{0.08}$
27	great$^{1.31}$ food$^{1.09}$ place$^{1.04}$ service$^{0.50}$ atmosphere$^{0.29}$ friendly$^{0.18}$ nice$^{0.16}$ looking$^{0.08}$ enjoy$^{0.06}$ fun$^{0.06}$
28	food$^{0.61}$ mexican$^{0.20}$ like$^{0.17}$ restaurants$^{0.14}$ restaurant$^{0.13}$ best$^{0.12}$ chips$^{0.10}$ area$^{0.10}$ order$^{0.09}$ just$^{0.09}$
29	salad$^{0.42}$ sauce$^{0.22}$ served$^{0.22}$ bread$^{0.15}$ cheese$^{0.13}$ pasta$^{0.12}$ garlic$^{0.10}$ italian$^{0.08}$ chicken$^{0.07}$ green$^{0.07}$

Tabela 28: Tópicos estimados com o modelo **Nota-Aspecto LDA** e **50 tópicos** (5 avaliativos e 40 não avaliativos) no conjunto de dados we8there. As dez palavras de maior pontuação são mostradas em ordem descrescente de pontuação para cada tópico.

Tópico	10 Palavras mais pontuadas
comida-	tasted$^{0.20}$ bland$^{0.18}$ didn't$^{0.16}$ sushi$^{0.13}$ dry$^{0.11}$ looked$^{0.10}$ overcooked$^{0.10}$ tasteless$^{0.09}$ better$^{0.09}$ mediocre$^{0.08}$
comida+	sushi$^{0.57}$ fresh$^{0.28}$ chef$^{0.25}$ roll$^{0.16}$ incredible$^{0.14}$ delicious$^{0.09}$ japanese$^{0.09}$ spicy$^{0.09}$ dishes$^{0.08}$ rolls$^{0.08}$
serviço-	minutes$^{0.56}$ waiter$^{0.21}$ took$^{0.19}$ waited$^{0.16}$ asked$^{0.15}$ order$^{0.15}$ finally$^{0.14}$ manager$^{0.13}$ arrived$^{0.10}$ wanted$^{0.10}$
serviço+	excellent$^{0.88}$ highly$^{0.37}$ recommend$^{0.35}$ atmosphere$^{0.32}$ delicious$^{0.31}$ outstanding$^{0.18}$ wonderful$^{0.15}$ friendly$^{0.14}$ definitely$^{0.13}$ cozy$^{0.11}$
valor-	said$^{0.41}$ did$^{0.35}$ asked$^{0.30}$ told$^{0.29}$ manager$^{0.29}$ came$^{0.12}$ poor$^{0.12}$ left$^{0.09}$ tip$^{0.07}$ charged$^{0.07}$
valor+	best$^{0.85}$ love$^{0.52}$ favorite$^{0.37}$ try$^{0.34}$ excellent$^{0.28}$ wonderful$^{0.16}$ perfect$^{0.15}$ italian$^{0.12}$ week$^{0.11}$ absolutely$^{0.10}$
ambiente-	owner$^{0.27}$ customers$^{0.16}$ rude$^{0.13}$ card$^{0.12}$ credit$^{0.12}$ charge$^{0.11}$ instead$^{0.09}$ problem$^{0.08}$ wrong$^{0.07}$ customer$^{0.07}$
ambiente+	beautiful$^{0.21}$ traditional$^{0.09}$ tea$^{0.08}$ romantic$^{0.08}$ decor$^{0.08}$ world$^{0.08}$ dining$^{0.06}$ outdoor$^{0.06}$ selection$^{0.06}$ superb$^{0.06}$
experiência-	bad$^{0.50}$ terrible$^{0.28}$ rude$^{0.21}$ cold$^{0.19}$ horrible$^{0.18}$ ok$^{0.17}$ waitress$^{0.14}$ awful$^{0.14}$ left$^{0.13}$ worst$^{0.10}$
experiência+	great$^{2.79}$ wonderful$^{0.39}$ atmosphere$^{0.30}$ fantastic$^{0.26}$ awesome$^{0.22}$ amazing$^{0.16}$ place$^{0.13}$ loved$^{0.11}$ friendly$^{0.10}$ outstanding$^{0.05}$
10	shrimp$^{0.47}$ seafood$^{0.45}$ crab$^{0.40}$ fish$^{0.32}$ fresh$^{0.27}$ san$^{0.17}$ salmon$^{0.11}$ lobster$^{0.10}$ bay$^{0.09}$ francisco$^{0.08}$
11	served$^{0.23}$ sauce$^{0.19}$ mushrooms$^{0.09}$ garlic$^{0.08}$ potatoes$^{0.08}$ entrees$^{0.08}$ grilled$^{0.08}$ cheese$^{0.07}$ topped$^{0.06}$ include$^{0.05}$
12	good$^{1.98}$ nice$^{0.84}$ food$^{0.56}$ quite$^{0.25}$ little$^{0.14}$ priced$^{0.14}$ tasty$^{0.10}$ atmosphere$^{0.10}$ reasonably$^{0.07}$ clean$^{0.06}$
13	salad$^{0.66}$ bread$^{0.45}$ italian$^{0.25}$ pasta$^{0.23}$ chicken$^{0.17}$ served$^{0.17}$ soup$^{0.16}$ dressing$^{0.12}$ fresh$^{0.09}$ sauce$^{0.09}$
14	sandwich$^{0.18}$ beef$^{0.17}$ lunch$^{0.14}$ got$^{0.14}$ pretty$^{0.12}$ quick$^{0.10}$ shop$^{0.10}$ downtown$^{0.09}$ deli$^{0.09}$ counter$^{0.08}$
15	want$^{0.42}$ just$^{0.34}$ people$^{0.33}$ eat$^{0.22}$ make$^{0.20}$ order$^{0.13}$ like$^{0.13}$ taste$^{0.12}$ way$^{0.11}$ come$^{0.11}$
16	mexican$^{0.36}$ chips$^{0.22}$ food$^{0.16}$ chicken$^{0.15}$ like$^{0.14}$ salsa$^{0.14}$ authentic$^{0.09}$ tacos$^{0.08}$ taco$^{0.07}$ order$^{0.07}$
17	dishes$^{0.41}$ chicken$^{0.37}$ chinese$^{0.20}$ rice$^{0.20}$ dish$^{0.18}$ thai$^{0.14}$ lunch$^{0.13}$ restaurant$^{0.12}$ fried$^{0.11}$ soup$^{0.11}$
18	restaurant$^{1.47}$ family$^{0.84}$ food$^{0.30}$ service$^{0.15}$ eat$^{0.14}$ children$^{0.12}$ kids$^{0.12}$ owned$^{0.10}$ restaurants$^{0.09}$ friendly$^{0.07}$
19	new$^{0.85}$ years$^{0.54}$ year$^{0.20}$ old$^{0.19}$ visit$^{0.14}$ ago$^{0.14}$ used$^{0.10}$ location$^{0.10}$ city$^{0.10}$ york$^{0.09}$
20	ordered$^{0.78}$ husband$^{0.45}$ did$^{0.26}$ server$^{0.23}$ waiter$^{0.16}$ said$^{0.14}$ entree$^{0.13}$ appetizer$^{0.12}$ dinner$^{0.11}$ delicious$^{0.10}$
21	good$^{1.18}$ really$^{0.86}$ like$^{0.47}$ little$^{0.30}$ bad$^{0.30}$ places$^{0.26}$ pretty$^{0.26}$ place$^{0.26}$ better$^{0.24}$ lot$^{0.12}$
22	got$^{0.31}$ really$^{0.21}$ little$^{0.17}$ didn't$^{0.17}$ came$^{0.12}$ wasn't$^{0.12}$ just$^{0.11}$ said$^{0.10}$ j$^{0.10}$ it's$^{0.10}$
23	room$^{0.65}$ dining$^{0.39}$ tables$^{0.16}$ right$^{0.11}$ main$^{0.10}$ near$^{0.08}$ bar$^{0.08}$ building$^{0.08}$ huge$^{0.08}$ walls$^{0.08}$
24	experience$^{1.37}$ dining$^{0.58}$ restaurant$^{0.50}$ wife$^{0.35}$ pleasant$^{0.21}$ service$^{0.19}$ fine$^{0.16}$ enjoyed$^{0.16}$ return$^{0.10}$ wonderful$^{0.08}$
25	great$^{1.68}$ place$^{1.36}$ food$^{0.92}$ service$^{0.22}$ love$^{0.13}$ gone$^{0.10}$ vegas$^{0.09}$ eat$^{0.08}$ atmosphere$^{0.07}$ waitresses$^{0.06}$
26	buffet$^{0.51}$ breakfast$^{0.47}$ diner$^{0.18}$ lunch$^{0.16}$ hotel$^{0.14}$ sunday$^{0.14}$ brunch$^{0.13}$ eggs$^{0.13}$ toast$^{0.09}$ selection$^{0.09}$
27	best$^{1.39}$ area$^{0.37}$ town$^{0.28}$ i've$^{0.27}$ try$^{0.23}$ real$^{0.19}$ better$^{0.15}$ far$^{0.12}$ restaurants$^{0.12}$ serve$^{0.12}$
28	pizza$^{1.22}$ cheese$^{0.34}$ chicago$^{0.22}$ order$^{0.15}$ crust$^{0.13}$ style$^{0.13}$ good$^{0.11}$ chain$^{0.10}$ pizzas$^{0.10}$ hungry$^{0.08}$
29	bbq$^{0.40}$ sauce$^{0.35}$ pork$^{0.28}$ ribs$^{0.21}$ sweet$^{0.12}$ meat$^{0.11}$ chicken$^{0.11}$ beans$^{0.10}$ flavor$^{0.09}$ potato$^{0.09}$
30	cafe$^{0.34}$ local$^{0.30}$ outside$^{0.22}$ inside$^{0.16}$ day$^{0.11}$ coffee$^{0.11}$ seating$^{0.10}$ number$^{0.09}$ park$^{0.09}$ city$^{0.08}$
31	wait$^{0.46}$ seated$^{0.22}$ table$^{0.21}$ 2$^{0.19}$ minutes$^{0.19}$ hour$^{0.17}$ arrived$^{0.17}$ people$^{0.17}$ reservations$^{0.16}$ busy$^{0.14}$
32	steak$^{0.60}$ cooked$^{0.20}$ dinner$^{0.17}$ special$^{0.16}$ rib$^{0.14}$ house$^{0.13}$ meal$^{0.11}$ steaks$^{0.11}$ prime$^{0.11}$ price$^{0.11}$
33	menu$^{0.92}$ variety$^{0.24}$ large$^{0.20}$ items$^{0.19}$ lunch$^{0.14}$ salads$^{0.14}$ available$^{0.14}$ offers$^{0.13}$ located$^{0.12}$ specials$^{0.12}$
34	food$^{2.06}$ service$^{1.58}$ excellent$^{0.34}$ return$^{0.22}$ better$^{0.12}$ slow$^{0.11}$ atmosphere$^{0.09}$ fast$^{0.09}$ wonderful$^{0.06}$ overall$^{0.06}$
35	restaurant$^{0.43}$ customers$^{0.15}$ care$^{0.15}$ review$^{0.14}$ business$^{0.13}$ clean$^{0.12}$ location$^{0.12}$ establishment$^{0.11}$ servers$^{0.10}$ needs$^{0.09}$
36	wine$^{0.77}$ list$^{0.27}$ glass$^{0.17}$ course$^{0.16}$ chef$^{0.10}$ restaurant$^{0.10}$ duck$^{0.09}$ bottle$^{0.08}$ menu$^{0.08}$ fine$^{0.08}$
37	dessert$^{0.43}$ cream$^{0.36}$ chocolate$^{0.34}$ ice$^{0.26}$ cake$^{0.17}$ hot$^{0.10}$ cold$^{0.08}$ butter$^{0.08}$ taste$^{0.08}$ delicious$^{0.08}$
38	food$^{0.75}$ times$^{0.49}$ visit$^{0.42}$ restaurant$^{0.42}$ eaten$^{0.32}$ recommend$^{0.28}$ time$^{0.27}$ 3$^{0.19}$ service$^{0.13}$ highly$^{0.10}$
39	it's$^{1.29}$ place$^{0.76}$ eat$^{0.29}$ you're$^{0.26}$ i've$^{0.18}$ i'm$^{0.16}$ isn't$^{0.12}$ you'll$^{0.10}$ looking$^{0.09}$ expect$^{0.09}$
40	time$^{0.55}$ make$^{0.44}$ home$^{0.37}$ going$^{0.26}$ way$^{0.23}$ eat$^{0.21}$ stop$^{0.18}$ worth$^{0.16}$ long$^{0.16}$ sure$^{0.16}$
41	waitress$^{0.51}$ came$^{0.36}$ table$^{0.32}$ check$^{0.21}$ drink$^{0.21}$ order$^{0.21}$ brought$^{0.20}$ took$^{0.16}$ meal$^{0.14}$ did$^{0.13}$
42	night$^{0.52}$ dinner$^{0.41}$ went$^{0.40}$ time$^{0.37}$ going$^{0.21}$ friend$^{0.18}$ tell$^{0.15}$ saturday$^{0.13}$ decided$^{0.13}$ called$^{0.11}$
43	service$^{0.39}$ special$^{0.30}$ friends$^{0.28}$ evening$^{0.24}$ enjoy$^{0.17}$ happy$^{0.16}$ enjoyed$^{0.14}$ dinner$^{0.14}$ party$^{0.13}$ meal$^{0.13}$
44	fries$^{0.42}$ burger$^{0.39}$ sandwich$^{0.33}$ cheese$^{0.22}$ chicken$^{0.12}$ burgers$^{0.10}$ bacon$^{0.10}$ menu$^{0.09}$ onion$^{0.09}$ ordered$^{0.08}$
45	prices$^{0.65}$ price$^{0.46}$ portions$^{0.38}$ quality$^{0.34}$ reasonable$^{0.32}$ small$^{0.28}$ food$^{0.26}$ worth$^{0.20}$ high$^{0.16}$ large$^{0.15}$
46	staff$^{1.17}$ friendly$^{0.80}$ food$^{0.67}$ wait$^{0.31}$ atmosphere$^{0.31}$ attentive$^{0.17}$ pleasant$^{0.14}$ service$^{0.13}$ decor$^{0.12}$ area$^{0.10}$
47	like$^{0.56}$ good$^{0.29}$ hot$^{0.24}$ just$^{0.19}$ people$^{0.18}$ wings$^{0.18}$ think$^{0.13}$ table$^{0.13}$ don't$^{0.13}$ right$^{0.10}$
48	know$^{0.55}$ don't$^{0.39}$ didn't$^{0.31}$ say$^{0.30}$ think$^{0.26}$ just$^{0.24}$ like$^{0.14}$ way$^{0.11}$ let$^{0.11}$ plate$^{0.10}$
49	bar$^{1.16}$ place$^{0.38}$ beer$^{0.20}$ drinks$^{0.17}$ pub$^{0.13}$ music$^{0.13}$ area$^{0.13}$ great$^{0.11}$ good$^{0.10}$ nice$^{0.10}$

ANEXO B – REPRODUÇÃO DO ARTIGO REPRESENTING AND CLASSIFYING USER REVIEWS

Representing and Classifying User Reviews

Denis D. Mauá[1], **Fabio G. Cozman**[1]

[1]Escola Politécnica – Universidade de São Paulo (USP)
CEP 05508-900 – São Paulo – SP – Brazil

{denis.maua,fgcozman}@usp.br

Abstract. *A large number of user reviews in the internet contains valuable information on services and products; for this reason, there is interest in automatically understanding such reviews. Sentiment Classification labels documents according to the feelings they express; instead of classifying a document into topics (sports, economics, etc), one attempts to tag the document according to overall feelings. Compared to the accuracy of traditional text categorization methods, sentiment classifiers have shown poor performance. We argue that such bad results are due to an improper representation of reviews. We describe a weakly supervised method that converts raw text into an appropriate representation, and show how techniques from information retrieval can acquire labeled data and process data using Markov logic. We report results on sentence classification and rating prediction that support our claims.*

1. Introduction

Review sites such as *Yelp (http://www.yelp.com)* and *Amazon (http://www.amazon.com)* encourage users to post reviews describing their opinions on products and services. These opinions can be used by other users to make informed decisions concerning purchases. Businesses can take advantage of reviews by obtaining consumer feedback on their products and services. Despite these potential benefits, the overwhelming number of reviews leads to an information overload that prevents users from fully exploiting the data.

Previous work on mining opinions from reviews have tried to summarize a document by its overall sentiment as a way to avoid information overload [Pang et al. 2002, Turney 2001, Pang 2005]. The focus has been on classification of texts according to sentiment indicators that can be binary or categorized. When compared to traditional text categorization [Manning and Schütze 1999], work on sentiment classification have reported poor classification performance.

We argue that these poor results are due to the lack of a proper representation for reviews. We propose a new representation that is better suited to such data (Section 3), and show how to obtain the representation from raw text using a weakly supervised method (Section 4). To do so, we use information retrieval techniques to acquire labeled data and Markov logic to specify classifiers [Richardson and Domingos 2006]. We report results for sentence classification and review rating prediction in Section 5.

2. Background

Information Retrieval (IR) techniques aim to retrieve relevant information from a large collection of (text) documents. The first step of an IR system is to index all documents by

building a *co-occurrence matrix*. Let $\{D_i\}$ be a collection of N documents and F a M-sized set containing the vocabulary for collection $\{D_i\}$, i.e. the set of all distinct words appearing in some document. Hence, the co-occurrence matrix C is a $M \times N$ matrix, whose rows represent term distributions and columns represent documents in the bag-of-words representation. Each cell $C_{i,j}$ is equal the number of occurrences of the ith term in F in the jth document of the collection. A heuristic that has been reported to improve the accuracy of document retrieval is **tf-idf** weighting that takes an element of matrix C to be $C_{i,j} = tf_{i,j} \times idf_i$, where tf is the term frequency given by $tf_{i,j} = C_{i,j} / \sum_i C_{i,j}$, and idf is the inverse document frequency given by $idf_i = \log N - \log \sum_j C_{i,j}$.

Documents are retrieved by means of a *query* document. A query is simply a vector of length M where the terms in F that we want to search for are set to one. Documents are then ranked by their similarity to the query. The most common similarity metric used in the *cosine* given by $sim(v_1, v_2) = (v_1 \cdot v_2)/(|v_1||v_2|)$, where v_1 and v_2 are two documents in a proper vector representation, and $|v|$ denotes the norm of a vector v.

A difficulty in comparing documents with the cosine metric is that it matches words that co-occur in two documents, but many different words can be used to describe a same target-information. Also, words often have several meanings, and simple term matching may lead to the overestimation of document relevance [Manning and Schütze 1999]. One way to overcome these problems is to include *term co-occurrence* information in the document representations. If a word w_1 co-occurs often with another word w_2, they are likely to share some relation. Thus a query for w_1 might include documents where w_2 appear and vice-versa. A reasonable way to accomplish this is by producing a *low-rank approximation* of matrix C. The new canonical terms of the matrix can then be understood as latent concepts that are able to generalize the original high dimensional term vector F to a new lower dimensional vector F' of concepts. This technique of dimensionality reduction is known in the IR literature by the name of **latent semantic indexing** (LSI), which contrasts the usual word indexing by the new latent concept indexing . The k-rank LSI representation C_k of co-occurrence matrix C is given by $C_k = U\Sigma_k V$, where U and V are the term and document matrices, respectively, given by the *single value decomposition* of matrix C, and Σ_k is the diagonal matrix of the k greatest singular values of C.

In this paper we build classifiers based on **Markov logic** (ML). Markov Logic is a statistical relational language that uses a first-order logic (FOL) syntax to specify complex Markov networks [Richardson and Domingos 2006]. Formally, a knowledge base (KB) in Markov logic is a set of (implicitly conjoined) weighted first-order formulae. Let x_i be a ground atom with truth value assigned (e.g. $\texttt{HasWord}(\texttt{D}, \texttt{"meal"}) : True$) and $x = \{x_i\}$ an interpretation (i.e., the set of all possible ground atoms with truth values assigned). Then, the probability of a particular interpretation is given by

$$P(x) = \frac{1}{Z} \exp\left(\sum_i w_i f_i\right), \tag{1}$$

where Z is the partition function given by $\sum_{x \in X} \exp\left(\sum_i w_i f_i\right)$, and w_i denotes the weight attached to the ith grounded formula f_i.

For instance, a *MaxEnt* Text Classifier [Manning and Schütze 1999] can be imple-

Review 1. The service was great but the food was terrible.

Review 2. The food was great but service was terrible.

Figure 1. Examples of user reviews.

mented in Markov Logic by the following model.

$$\begin{array}{rl} w_{\mathtt{w+,a+}} & \mathtt{HasWord(s,w+)} \rightarrow \mathtt{Topic(s,a+)} \\ w_{\mathtt{a+}} & \neg\mathtt{Topic(s,a+)} \end{array}$$

The first formula models how evidences (the presence of a word in a sentence) collaborate to topic discrimination under a naive Bayes assumption that word contributions are independent. The second formula imposes a prior on topic distributions. It represents that in the absence of any evidence no aspect topic should be true. The $+$ marks are syntactic sugars that indicate that these formulas are actually templates. In practice, if W is the number of words in the vocabulary and K the number of topics, then the first formula is turned into $W * K$ formulas, and the second formula becomes K prior formulas, one for each topic. The $(W + 1)K$ weights are then discriminatively learned from data.

For example, a document `D` with only word *meal* occurring has probabilities

$$\begin{array}{l} P(\mathtt{Topic(D,"food")}|\mathtt{HasWord(D,"meal")}) \propto \\ P(\mathtt{Topic(D,"food")},\mathtt{HasWord(D,"meal")}) = \frac{1}{Z}\exp(w_{\mathtt{meal,food}}) \end{array}$$

of belonging to topic *food* and $P(\neg\mathtt{Topic(D,"food")}|\mathtt{HasWord(D,"meal")}) \propto \frac{1}{Z}\exp(w_{\mathtt{food}})$ of not belonging to, given by Equation (1).

A considerable advantage of Markov Logic modeling is the availability of efficient methods for learning and inference and an opensource implementation (available at http://alchemy.cs.washington.edu).

3. A Proper Representation for Reviews

Sentiment classification methods usually assume that a review can be summarized by a single overall metric. However, as noted in [Snyder and Barzilay 2007] and [Titov and McDonald 2008], opinions expressed by reviewers are multi-faceted and cannot be correctly represented by a single sentiment score. Take the example reviews in Figure 1. Both examples express opposite sentiments on different aspects of an object. A neutral sentiment might be assigned to both reviews which clearly does not represent them well. A more reasonable assumption is that reviews can be summarized by aspect-based sentiments. For instance, we can classify reviews in Figure 1 as being, respectively, positive and negative according to aspect *service* and negative and positive, respectively, according to aspect *food*. We call the task of classifying documents according to the sentiments they express regarding a particular aspect as *aspect-based sentiment classification*.

Machine learning sentiment classification methods use the traditional *bag-of-words* model to represent reviews. Such a representation assumes that each document can be represented in the vector space by a function of the number of occurrences of each word in text. A nice way to visualize this representation is to see each document as an unordered list of words; Figure 2 shows the representation for both documents in Figure 1.

service food great terrible was but

Figure 2. Bag-of-words representation for *both* documents in Figure 1.

Review 1

service/*service* food/*food* great/*service* horrible/*food* was/*service* was/*food* but/*other*

Review 2

service/*service* food/*food* great/*food* horrible/*service* was/*service* was/*food* but/*other*

Figure 3. A proper representation for documents in Figure 1. Documents are no longer mapped to the same representation as in Figure 2.

The bag-of-words representation assumes that documents regarding different subjects have different word distributions. For example, in sport articles one should expect to find more occurrences of word "athlete" than in economic articles. Thus, we can predict the topic of a document by merely checking the word occurrences in it. However, this assumption fails when modeling reviews to sentiment classification, because here we are not concerned with the main subject of the review (e.g. restaurants, hotels or electronics) but with the many "micro"-opinions in the text. In fact, one can conceive all reviews as belonging to one same class, the class of reviews, thus expecting all of them to present similar word distributions. This explains the poor performance of sentiment classifiers based on the bag-of-words model.

A better representation for aspect-based sentiment classification is then to assign a tag to each word so as to discriminate with respect to the aspect it refers to. Figure 3 depicts a possible representation of reviews in Figure 1 according to this new model. Note that unlike the bag-of-words representation depicted in Figure 2 this new representation disambiguates between the two documents. A *bag-of-tagged-words* model is a representation model for text documents that extends the bag-of-words model as follows. Let $D = \{w_1, \ldots, w_N\}$ denote a document, where w_i denotes the ith token (word) of the document and N the length of the document. Let also $A = \{a_1, \ldots, a_M\}$ denote the set of topics, where each a_i denotes a distinct topic that a word may refer to.

Definition 1 *The* bag-of-tagged-words *for document D is the* bag-of-words *model for the new document $\bar{D} = \{t_1, \ldots, t_N\}$, where $t_i = (w_i, \{a_j\})$ is an ordered pair of the ith token and a subset $\{a_j\} \in A$ of the topic set.*

The main goal of this work is to develop a method that, given a set of topics A and a document D, generates a new document $\bar{D}$ represented as a bag-of-tagged-words. For reviews, the set of topics A is the set of aspects of an object that a user can comment on. Note that this definition of a bag-of-tagged-words model is a broad definition which fits any text document, not only reviews. In fact, this model can be seen as an instance of the general class of *topic models*, used in information retrieval and NLP to represent text documents [Manning and Schütze 1999].

4. The Method

An overview of the proposed method is depicted in Figure 4. Given a collection of documents D of size N and set of aspects A of size K, the system first produces an indexing of all sentences in the collection. Then, for each aspect $a \in A$ it retrieves the top-k most

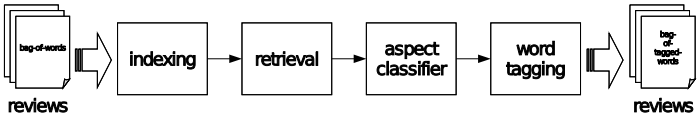


Figure 4. Method Overview. The diagram illustrates the main steps in the process of converting text documents to the *bag-of-tagged-words* representation.

Table 1. Aspect word sets used. Right column shows queries that retrieve sentences for aspects on the left column.

food	food dish course meal chicken portion taste soup bread pasta salad meat chicken
service	service staff wait waiter hostess host
value	price value cost worth cheap expensive
atmosphere	atmosphere ambiance ambience decor crowded noisy loud comfortable
experience	experience overall place favorite visit fun

relevant sentences to a and produces a labeled set by assigning label a to the k sentences retrieved for a. This training dataset is used to learn a Markov logic classifier capable of categorizing sentences according to the aspect they refer to. The classifier is then used to build a new *bag-of-tagged-words* representation for documents in D by classifying each sentence in $d \in D$ according to the aspects $a \in A$ and by labeling each word in D by the predicted label(s) of the sentences they are in. The result is a new collection $\bar{D}$ where documents follow the new representation scheme. The next subsections detail each step.

4.1. Indexing

The first step is to build an index of all sentences in the collection. First, all documents are segmented into sentences, forming a new collection S of all sentences of all documents. Because reviews are very noisy,[1] a simple split-by-period procedure yielded best results in sentence segmentation than statistical methods. With each document segmented into sentences a co-occurrence matrix C_k of all sentences is built, followed by the application of tf-idf weighting and latent semantic indexing .

4.2. Retrieval

After indexing the sentences, the next step is to retrieve relevant sentences to compose a training dataset. In order to find sentences that are relevant to a given aspect we need to construct appropriate queries. This is very subjective and is the only part of our method where human supervision is required. An intuitive strategy is to look at text corpora and extract the most common words used by reviewers when commenting on a given aspect.

To facilitate the extraction of relevant words we have designed a bigram filter that returned the most frequent bigrams (pairs of adjacent words) whose first word is an adjective and second word is a noun. Although many other patterns could be investigated, we found this heuristic very satisfactory. We produced a list of approximately 50 words (adjectives and nouns) and manually cluster them into K groups, each group representing a distinct aspect in A. Table 1 shows the set of words used in our experiments with

[1] By noisy, we mean that simple formal rules such as capitalizing initial words and separating sentences with spaces are often not followed by review authors.

restaurant reviews. The small number of words used to represent each aspect is only feasible because of the LSI transformation applied to the co-occurrence sentence matrix, which makes other non-listed related terms relevant to the query.

With a set of relevant words for each aspect in hand, we are able to build queries for retrieving relevant sentences. For each aspect $a \in A$ the top-k sentences more similar for the query q_a of relevant words of a are retrieved and labeled as a. The final result is a $(K \times k)$-length training dataset with equally distributed classes, which we can use to learn a sentence classifier.

4.3. Aspect Classifier

In this step we learn a K-topic MaxEnt classifier in Markov Logic, which classifies each sentence as whether it belongs to each of the aspects $a \in A$ as described in Section 2.

4.4. Word Tagging

The last step of the method is to tag each token in a document with aspects so as to produce a new *bag-of-tagged-words* representation. The process is described in pseudo-code by Algorithm 1. For each document in the collection, we segment it into sentences as explained in Subsection 4.1 (line 2). Then, a *bag* structure is created to store the new representation (line 3). Each sentence in the original document is classified with respect to each aspect as follows. For each aspect, a binary classifier is run to predict whether the sentence belongs to that aspect (lines 5–9). If the output is positive, every word in the sentence is tagged with the aspect a tag and added to the bag. If not, words are added with no tag. A untagged word can be fitted in our definition of a bag-of-tagged-words (Definition 1) by adding a dummy extra topic to the aspect set and tagging untagged words with this new topic. Finally, the new bag is stored in the collection.

```
1  foreach document in the collection do
2      segment document into sentences
3      create new bag
4      foreach sentence in document do
5          foreach aspect in A do
6              if Classify (sentence,aspect) then
7                  foreach word w ∈ S do add word/aspect to bag
8              else
9                  add word to bag
10     document ← bag
```

Algorithm 1: Word Tagging. Given a collection of documents, algorithm generates a collection of documents as a *bag-of-tagged-words*.

5. Experiments

We evaluate the proposed method using 6260 restaurant reviews downloaded from the we8there website (http://www.we8there.com). Each review is composed by a short text (on average 90 words) and a set of five ratings on a 1–5 scale regarding aspects *food*, *service*, *value*, *atmosphere* and *overall experience*. We report results on the sentence classification and aspect-based rating prediction tasks.

Table 2. Results for the sentence classification task with different indexing schemes. The numbers report on F1-measure (in %).

Representation	Food	Service	Value	Atmosphere	Experience	**Average**
COUNT	47.80	40.36	**31.82**	19.36	24.72	32.81
TF-IDF	34.30	33.33	15.62	23.33	37.17	28.75
COUNT LSI	**47.84**	**67.53**	21.18	35.51	40.00	42.41
TF-IDF LSI	44.22	61.87	28.92	**39.37**	**40.31**	**42.94**

5.1. Sentence Classification

In this task, we evaluate the performance of the method in extracting useful sentences for training a Markov Logic classifier as well as the classification performance. The objective is for each aspect to classify a sentence as whether it comments on the aspect. We segmented all reviews into sentences in the dataset resulting in a database of 49662 sentences. We then filtered each sentence by removing low-frequency and function words, ending with a 3402x49662 word-sentence co-occurrence matrix. We extracted 500 sentences from this dataset and manually labeled each sentence, so that each had from zero up to five labels. Then, all sentences were converted into binary vectors: for each sentence i of the dataset a 3402-length vector was created by assigning 1 to the jth position iff term t_j occurs in i, and 0 otherwise. This binary representation helps improve classification accuracy. We use F1-measure to evaluate classification. F1-measure is the harmonic mean of precision and recall, given by F1 $= 0.5(p \times r)/(p + r)$, where p is the precision given by $p =$ (# of correct classified instances)/(# of total classified instances) and r is the recall given by $r =$ (# of correct classified instances)/(# of total instances belonging to aspect).

Evaluating Indexing Schemes In order to assess the different indexing schemes, we selected the top-500 sentences more relevant according to each aspect, ending with a 2500-length training dataset. Because a sentence may be relevant to more than one aspect, some sentences in the training data may occur more than once. Table 2 presents the results for different schemes according to the F1-measure, with best results for each aspect in bold. The COUNT scheme is the traditional bag-of-words model. The TF-IDF is the result of the tf-idf weighting to this vector. COUNT LSI and TF-IDF LSI are the COUNT and TF-IDF matrices, respectively, after latent semantic indexing being applied. On average, the TF-IDF LSI performed better than others, indicated by its higher score in the last column. Surprisingly, the COUNT scheme had the best result for aspect *value*. A possible explanation is that aspect *value* can actually be regarded as a sub-aspect of aspect *food*. This way, LSI schemes may increase the ratio of noisy by adding many *food* documents to the *value* set, hurting classifier accuracy.

Evaluating k Influence In order to assess the influence of the number of retrieved sentences per aspect in the training set, we performed an experiment with TF-IDF LSI indexing and k varying from 100 to 2000. As Table 3 shows, the best performance according to the average F1-measure happens when $k = 500$. For too small k, there is no sufficient number of training instances to correctly discriminate data. For too large k, the ratio of misclassified data in the training set increases, leading to poor performance.

Evaluating Classifiers We evaluated the performance of the Markov Logic Classifier against two baselines. The first baseline is a procedure that classifies all sentences as

Table 3. Results for the sentence classification task with varying k. Numbers report on F1-measure (in %).

k	Food	Service	Value	Atmosphere	Experience	**Average**
100	24.13	56.67	28.13	33.71	41.59	36.84
200	26.81	**65.65**	22.53	34.29	**41.91**	38.24
500	44.22	61.87	**28.92**	**39.37**	40.31	**42.94**
1000	**50.00**	57.33	25.00	37.41	32.47	40.44
2000	46.53	50.73	18.41	29.27	30.04	35.00

Table 4. Results for the sentence classification task with different classifiers. Numbers report on F1-measure (in %).

	Food	**Service**	**Value**	**Atmosphere**	**Experience**	**Average**
Baseline	**45.14**	15.50	5.04	12.80	13.05	18.31
Naive Bayes	38.06	43.55	13.56	34.29	32.43	32.38
Markov Logic	44.22	**61.87**	**28.92**	**39.37**	**40.31**	**42.94**

belonging to all aspects. Its precision is simply the number of sentences of each aspect over the total. For the second baseline we implemented a Naive Bayes Classifier. Naive Bayes Classifiers have been reported as perform well for text categorization [Manning and Schütze 1999]. Table 4 presents the results obtained with $k = 500$ and TF-IDF LSI indexing. The Baseline refers to the first baseline classifier. Except for aspect *food*, Markov Logic performs much better than the others. The poor performance for aspect *food* is due to a low recall. This is because most part of the sentences comment on food quality, thus simply classifying sentences as belonging to food leads to good results. In fact, Markov Logic had the highest precision in aspect food (92.86% against 84.62% of Naive Bayes and only 45.14% for the Baseline). On the recall, however, results were far less satisfactory (29.02% against 24.55% for NB and 45.14% for Baseline).

5.2. Aspect-Based Rating Prediction

Aspect-Based Rating Prediction is the task of classifying review according to an aspect in a given pre-defined scale. In our dataset, ratings vary from 1–5. We performed a 80/20 split on the data, ending with 5008 and 1252 instances, respectively, for the training and test sets. A baseline was obtained by training a *MaxEnt* Markov Logic classifier with vectors following the common bag-of-words representation, except that only the presence/absence was stored as information. Then, using the algorithm described in Subsection 4.4, Markov Logic classifiers, and $k = 500$ and TF-IDF LSI scheme for sentence classification, we produce a bag-of-tagged-words representation of the dataset. Results are shown in Table 5. On average and in almost all aspects, the classifier learned with the bag-of-tagged-words (B-O-T-W) dataset performed slightly better than that learned with the common bag-of-words (B-O-W). By comparing the results on aspect *value* with those on Table 4, one can see that the sentence classification for this aspect had the poorest performance, which may explain the bad results in rating prediction.

6. Related Work

Sentiment Classification The problem of sentiment classification has been treated often as a binary classification task, where the goal is to predict the overall polarity (positive

Table 5. Results for the rating prediction task with different document representations. The numbers report on overall accuracy (in %).

Model	Features	Food	Service	Value	Atmosphere	Experience	**Average**
B-O-W	3402	61.84	54.00	**52.40**	48.32	59.36	55.18
B-O-T-W	18760	**64.16**	**55.20**	51.60	**48.40**	**60.16**	**55.90**

or negative) of the document. [Pang et al. 2002] provides a detailed analysis of machine learning methods to this task, and reports performance much lower than traditional topic classification ($< 85\%$ for sentiment classification against $> 95\%$ for general text categorization). They conclude that the *bag-of-words* model is one of the main factors of the low accuracy of the classifiers. In [Turney 2001], the author tries to avoid possible limitations of machine learning methods by using mutual information metrics and representing the overall sentiment as the average sentiment of the sentences. He uses a simple PMI estimator which scores each sentence with discriminative words such as "good" or "bad", reporting accuracies in the range $66\% - 84\%$. More recently, [Pang 2005] have allowed a finer-grained description of sentiments. They classify documents in four-classes (0–3) indicating the strength of sentiment, and using SVMs [Joachims 1998] they achieve maximum accuracies of $\sim 66\%$ for a movie review dataset. In [Snyder and Barzilay 2007], this model is extended to allow sentiments to be classified regarding different aspects. They use perceptron-like algorithms and meta-classifiers to include aspect sentiment correlation into their final classifier. They show that explicitly modeling aspect correlation improves performance of aspect-based rating predictors, but do not report on accuracy (they use rank loss to evaluate their method).

Opinion Extraction Research in Opinion Extraction tries to extract passages from text representatives of an opinion, and then group extracted passages by the fine-grained aspect they refer to. This way, a review can be seen as a series of polarized opinions on fine-grained aspects of the assessed object. [Hu and Liu 2004] extract frequent nouns from text by applying an associative rule learner and use a syntactic parser to check for possible modifiers (adjectives) for each extracted word within a sentence. They use a thesaurus based procedure to find the polarity of modifiers, and assign each sentence to the fine-grained (noun) aspect with a polarity tag. A review is then represented by a set of fine-grained (noun) aspects with opinion polarities attached. For example, the reviews 1 and 2 of Figure 1 would be represented by the vectors [`service+`,`food-`] and [`service-`,`food+`]. [Popescu and Etzioni 2005] use PMI estimation, with a priori knowledge of the world, to search for possible candidates to fine-grained aspects in text. They apply relaxation labeling, a common technique from image processing to determine the polarity of aspect noun modifiers (which they extend to adverbs and verbs). They report gains up to 11% with respect to previous work.

Topic Modeling Topic Modeling is concerned with richer representations of text documents. A topic model is a probabilistic model that jointly assign topic distributions to documents and words, and can be seen as a probabilistic version of the more common latent semantic indexing (LSI) method [Manning and Schütze 1999]. Perhaps the work closest to ours is [Titov and McDonald 2008], where the authors propose a probabilistic generative model to represent reviews. Unlike our work, their method is completely unsupervised. Their representation model is also close to ours, but in their model each word

is assigned a mixture model over the set of possible aspects. They report improvements on the review rating prediction task when compared to common bag-of-words model and also to other traditional Topic Models such as LDA and PLSA.

7. Conclusion and Future Work

Despite the success in traditional text categorization, machine learning methods have performed poorly on sentiment classification. Our claim is that such a poor performance is due to the lack of a proper representation of reviews. To support our claim, we have presented a novel representation for text documents that is better suited to sentiment classification. Our method relies on information retrieval techniques and Markov Logic to translate documents into this new representation with very little human intervention. We report results on aspect-based rating prediction showing that the proposed method indeed improves the performance.

In the future, we plan to investigate unsupervised mechanisms to automate the only step in this process where human intervention is necessary.

References

Hu, M. and Liu, B. (2004). Mining opinion features in customer reviews. In McGuinness, D. L. and Ferguson, G., editors, *AAAI*, Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA, pages 755–760. AAAI Press / The MIT Press.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. pages 137–142. Springer Verlag.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Pang, B. (2005). Seeing stars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL 05 ACL 05*, page 115.

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. Association for Computational Linguistics.

Popescu, A.-M. and Etzioni, O. (2005). Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT 05 HLT 05*, pages 339–346, Morristown, NJ, USA. Association for Computational Linguistics.

Richardson, M. and Domingos, P. (2006). Markov logic networks. *Machine Learning*, 62(1-2):107–136.

Snyder, B. and Barzilay, R. (2007). Multiple aspect ranking using the good grief algorithm. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 300–307, Rochester, New York. Association for Computational Linguistics.

Titov, I. and McDonald, R. (2008). Modeling online reviews with multi-grain topic models. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 111–120, New York, NY, USA. ACM.

Turney, P. D. (2001). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Morristown, NJ, USA. Association for Computational Linguistics.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo