UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

AUGUSTO CESAR SOUZA MARTINS

AVALIAÇÃO DE CARACTERÍSTICAS QUE INFLUENCIAM NOS VOTOS DE UTILIDADE DE OPINIÕES SOBRE SERVIÇOS EM PORTUGUÊS

DISSERTAÇÃO

CURITIBA

2015

AUGUSTO CESAR SOUZA MARTINS

AVALIAÇÃO DE CARACTERÍSTICAS QUE INFLUENCIAM NOS VOTOS DE UTILIDADE DE OPINIÕES SOBRE SERVIÇOS EM PORTUGUÊS

Dissertação apresentada ao Programa de Pós-graduação em Computação Aplicada da Universidade Tecnológica Federal do Paraná como requisito parcial para obtenção do grau de "Mestre em Computação Aplicada" – Área de Concentração: Engenharia de Sistemas Computacionais.

Orientador: Prof. Dr. Cesar Augusto Tacla

CURITIBA

2015

Dados Internacionais de Catalogação na Publicação

M386a
2015

Martins, Augusto Cesar Souza
Avaliação de características que influenciam nos votos de utilidade de opiniões sobre serviços em português / Augusto Cesar S. Martins.-- 2015.
83 p. : il. ; 30 cm

Dissertação (Mestrado) - Universidade Tecnológica Federal do Paraná. Programa de Pós-graduação em Computação Aplicada, Curitiba, 2015
Bibliografia: p. 68-71

1. Mineração de dados (Computação). 2. Opinião pública. 3. Hotéis - Avaliação. 4. Agentes de viagem - Avaliação. 5. Informática - Dissertações. I. Tacla, Cesar Augusto, orient. II. Universidade Tecnológica Federal do Paraná - Programa de Pós-graduação em Computação Aplicada. III. Título.

CDD: Ed. 22 -- 621.39

Biblioteca Central da UTFPR, Câmpus Curitiba

ATA DE DEFESA DE DISSERTAÇÃO DE MESTRADO Nº 34

Aos 27 dias do mês de agosto de 2015 realizou-se na sala B-204 a sessão pública de Defesa da Dissertação de Mestrado intitulada "Avaliação de Características que influenciam nos votos de utilidade de opiniões sobre serviços em Português", apresentada pelo aluno **Augusto Cesar Souza Martins** como requisito parcial para a obtenção do título de Mestre em Computação Aplicada, na área de concentração “Engenharia de Sistemas Computacionais”, linha de pesquisa “Sistemas Inteligentes e Lógica”.

Constituição da Banca Examinadora:

Prof. Dr. Cesar Augusto Tacla, UTFPR - CT (Presidente) _____

Prof. Dr. Celso Antonio Alves Kaestner, UTFPR – CT _____

Prof. Dr. Emerson Paraiso, PUC - PR _____

Prof. Dr. Laudelino Cordeiro Bastos, UTFPR – CT _____

Em conformidade com os regulamentos do Programa de Pós-Graduação em Computação aplicada e da Universidade Tecnológica Federal do Paraná, o trabalho apresentado foi considerado _____ (aprovado/reprovado) pela banca examinadora. No caso de aprovação, a mesma está condicionada ao cumprimento integral das exigências da banca examinadora, registradas no verso desta ata, da entrega da versão final da dissertação em conformidade com as normas da UTFPR e da entrega da documentação necessária à elaboração do diploma, em até _____ dias desta data.

Ciente (assinatura do aluno): _____

(para uso da coordenação)

A Coordenação do PPGCA/UTFPR declara que foram cumpridos todos os requisitos exigidos pelo programa para a obtenção do título de Mestre.

Curitiba PR, _____/_____/_____ _____

"A Ata de Defesa original está arquivada na Secretaria do PPGCA".

AGRADECIMENTOS

Agradeço a Deus por ter me dado força e saúde para concluir mais essa etapa da minha vida. A meus pais Luiz A. P. Martins e Shirley B. S. Martins pelo apoio e incentivo ao longo do caminho que trilhei até aqui. Aos meus irmãos Luiz Felipe e Ana Caroline e aos amigos Adão M. Ferreira, Luciano Arruda e Ricardo Venanti por também incentivar e torcer pelo sucesso. A minha namorada Mislene Sampaio pelo carinho, incentivo, compreensão e paciência nos momentos que mais precisei ao longo do mestrado.

Ao meu orientador, Prof. Cesar Augusto Tacla, pela disponibilidade, colaboração, dedicação e paciência com que me conduziu durante a realização deste trabalho. Posso afirmar que foi um privilégio tê-lo como meu orientador.

Aos professores Cesar A. Tacla, Celso A. A. Kaestner, Myriam R. B. S. Delgado e Laudelino C. Bastos pelos pertinentes apontamentos e sugestões realizados na apresentação deste projeto durante os seminários de acompanhamento desta pesquisa.

Aos professores Tania M. Centeno, Celso A. A. Kaestner, Laudelino C. Bastos, Adolfo G. S. S. Neto, Cesar A. Tacla, Murilo V. G. Silva, Myriam R. B. S. Delgado e Gustavo A. G. Lugo pelo relevantes conhecimentos transmitidos em suas disciplinas.

Aos professores Paulo H. Cayres, Douglas J. Peixoto e Margareth Poli principalmente pelo apoio e palavras de incentivo.

E finalmente, aos professores Cesar A. Tacla, Celso A. A. Kaestner, Laudelino C. Bastos e Emerson Paraíso pela participação na banca de defesa e pelos apontamentos realizados que contribuíram para a melhoria deste trabalho.

RESUMO

MARTINS, Augusto Cesar Souza. AVALIAÇÃO DE CARACTERÍSTICAS QUE INFLUENCIAM NOS VOTOS DE UTILIDADE DE OPINIÕES SOBRE SERVIÇOS EM PORTUGUÊS. 83 f. Dissertação – Programa de Pós-graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

O grande número de opiniões geradas por usuários *online* fez o antigo "boca a boca" migrar para o mundo virtual. Além de numerosas, muitas opiniões úteis estão misturadas com um grande número de opiniões fraudulentas, incompletas ou repetitivas. No entanto, como encontrar os fatores que influenciam no número de votos recebidos por uma opinião e encontrar opiniões consideradas úteis? A literatura na área de mineração de opiniões possui diversos estudos e técnicas que são capazes de analisar a influência de propriedades encontradas no texto das opiniões. Este trabalho apresenta a adaptação para o português de uma metodologia de avaliação de utilidade de opiniões com o objetivo de identificar quais características exercem maior influência na quantidade de votos de utilidade: básicas (ex. nota atribuída a produtos/serviços, data da publicação), textuais (ex. tamanho das palavras, parágrafos) e semântica (ex. o significado das palavras do texto). A avaliação foi realizada em uma base de dados extraída do *site* TripAdvisor com opiniões sobre hotéis escritas em português. Resultados mostram que os usuários dão mais atenção a opiniões recentes com notas mais altas para localização do hotel e com notas mais baixas para qualidade do sono, atendimento e limpeza. Textos com opiniões positivas, palavras curtas, poucos adjetivos e advérbios aumentam as chances de receber mais votos.

Palavras-chave: Mineração de opiniões, Utilidade da opinião, Qualidade da opinião

ABSTRACT

MARTINS, Augusto Cesar Souza. ASSESSEMENT OF FEATURES INFLUENCING THE VOTING FOR OPINIONS' HELPFULNESS ABOUT SERVICES IN PORTUGUESE. 83 f. Dissertação – Programa de Pós-graduação em Computação Aplicada, Universidade Tecnológica Federal do Paraná. Curitiba, 2015.

The large number of opinions generated by online users made the former "word of mouth" find its way to virtual world. In addition to be numerous, many of the useful reviews are mixed with a large number of fraudulent, incomplete or duplicate reviews. However, how to find the features that influence on the number of votes received by an opinion and find useful reviews? The literature on opinion mining has several studies and techniques that are able to analyze of properties found in the text of reviews. This paper presents the application of a methodology for evaluation of usefulness of opinions with the aim of identifying which characteristics have more influence on the amount of votes: basic utility (e.g. ratings about the product and/or service, date of publication), textual (e.g. size of words, paragraphs) and semantics (e.g., the meaning of the words of the text). The evaluation was performed in a database extracted from TripAdvisor with opinions about hotels written in Portuguese. Results show that users give more attention to recent opinions with higher scores for value and location of the hotel and with lowest scores for sleep quality and service and cleanliness. Texts with positive opinions, small words, few adjectives and adverbs increase the chances of receiving more votes.

Keywords: Opinions mining, Opinion usefulness, Opinion quality

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

AIC	*Akaike Information Criterion*
BOW	*Bag of Words*
SVD	*Singular Value Decomposition*
IMDB	*Internet Movie Database*
LSA	*Latent Semantic Analysis*
MCP	Modelo de Chances Proporcional
MR	*Misclassification Rate*
OLR	*Ordinal Logistic Regression*
PLN	Processamento de Linguagem Natural
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency-Inverse Document Frequency*

LISTA DE SÍMBOLOS

Σ Matriz diagonal

α Variáveis independentes do modelo de regressão

β Vetor de coeficientes de regressão

π Probabilidade de uma classe

SUMÁRIO

1 INTRODUÇÃO

O grande número de opiniões geradas por usuários em *blogs*, redes sociais, *micro-blogs*, lojas virtuais, *sites* especializados em avaliações de produtos e serviços, fez o antigo boca a boca migrar para o mundo virtual, criando grandes comunidades eletrônicas. O conteúdo gerado por usuários não possui uma estrutura formal e tampouco é simples de ser processado automaticamente (LIU, 2012).

No entanto, como o número de opiniões cresce rapidamente, é cada vez maior o esforço do usuário para encontrar a informação desejada. Além de numerosas, muitas das opiniões são fraudulentas, incompletas ou repetitivas. Complementar a isso, opiniões úteis estão misturadas com um grande número de opiniões sem utilidade, podendo sobrecarregar os usuários (PANG; LEE, 2005; LIU et al., 2008; LU et al., 2010; TSAPARAS et al., 2011).

Uma alternativa utilizada por empresas como Amazon e TripAdvisor foi implementar um sistema de avaliação onde seus usuários informam se a opinião de outra pessoa foi útil ou não. Essa avaliação se dá através da pergunta "Esta opinião foi útil para você?", geralmente posicionado na sequência do conteúdo da opinião e somente com as opções "Sim" ou "Não" como respostas.

A tarefa mais importante quando se organiza um conjunto de opiniões é determinar quais serão úteis para os usuários. Além da pergunta sobre a utilidade da opinião, o total de votos positivos e o total geral de votos também podem ser apresentados, sendo estes utilizados no cálculo do índice de utilidade da opinião. Este índice é usualmente um dos critérios de ordenação das opiniões dos *sites* que utilizam esse sistema de votação (GHOSE; IPEIROTIS, 2011).

Normalmente as opiniões são organizadas por data ou pontuação (i.e. nota ou números de estrelas atribuído pelo público) (KIM et al., 2006; TANG et al., 2013), porém, a maioria dos *sites* também apresenta um resumo da avaliação de todos que votaram, como "20 de 30 pessoas acham essa opinião útil" ou simplesmente o número de votos de utilidade que a opinião recebeu junto com o conteúdo do texto. No geral, as opiniões

são criadas por pessoas que possuem necessidades, pontos de vista e tiveram experiências diferentes sobre um mesmo item e revelam quais atributos do produto ou serviço são os mais importantes.

No entanto, o sistema de votação de utilidade não resolve todos os problemas, pois a maior parte das opiniões não recebe voto e sem isso, o mesmo não cumpre o seu papel de indicar o conteúdo mais útil (CAO et al., 2011). Na literatura, é possível encontrar algumas pesquisas com o objetivo de estimar automaticamente a qualidade das opiniões como Kim et al. (2006), Zhang e Varadarajan (2006), Ghose e Ipeirotis (2007), Liu et al. (2007), Tsur e Rappoport (2009), Liu et al. (2008) e Lu et al. (2010). Estas previsões de utilidade normalmente fazem uso de técnicas de processamento de texto como mineração de dados, aprendizado de máquina e processamento de linguagem natural (PLN) utilizando opiniões que receberam pelo menos um voto de utilidade.

Por outro lado, existem estudos como Talwar et al. (2007), Cao et al. (2011), Ghose e Ipeirotis (2011), Korfiatis et al. (2012), Lee (2013), Tang et al. (2013), que além de utilizarem as mesmas técnicas de processamento de texto citadas anteriormente, também analisam como as propriedades básicas (i.e. nota sobre o produto e/ou serviço, data da publicação), estilo do texto (i.e. tamanho das palavras, parágrafos, etc.), erros de escrita, semântica (i.e. o significado das palavras do texto), histórico de contribuições de um mesmo autor e as opiniões fortemente polarizadas (i.e. muito positivas ou negativas, de acordo com as notas atribuídas pelos usuários) influenciam na leitura e avaliação das opiniões publicadas.

Entre as propriedades encontradas nas opiniões, a "utilidade da opinião" representa uma atribuição de valor a uma avaliação subjetiva realizada por outra pessoa, além de agregar uma percepção de utilidade a informação contida no texto (CAO et al., 2011). Uma avaliação positiva útil agrega valor ao produto ou serviço, porém uma opinião negativa ou uma crítica pode ser uma oportunidade de corrigir falhas ou defeitos encontrados pelos clientes.

1.1 MOTIVAÇÃO

A abordagem do trabalho de Cao et al. (2011) se diferencia do restante da literatura, pois ao invés de tentar prever o nível de utilidade das opiniões que não receberam votos, o objetivo da pesquisa é determinar os fatores que influenciaram no número de

votos das opiniões por meio da análise de características básicas[1], estilo do texto[2] e características semânticas[3]. Nessa linha de pesquisa, Korfiatis et al. (2012) possuem trabalho semelhante a Cao et al. (2011) na tentativa de encontrar os aspectos que interferem na percepção de utilidade de uma opinião.

Korfiatis et al. (2012) utilizam a conformidade, ou seja, o quanto a nota da opinião sobre o produto avaliado está em acordo com a média geral das notas, métricas de inteligibilidade da língua inglesa (DUBAY, 2004) para verificar o nível de compreensão sobre a opinião e por último a influência do tamanho do texto da opinião, com o objetivo de determinar se essas características dos textos tem influência sobre o volume de vendas dos produtos analisados e se opiniões fáceis de ler tem mais chance de receber mais votos. Porém, os testes utilizados no artigo são específicos para a língua inglesa e não se aplicam à língua portuguesa (BARBOZA; NUNES, 2007; SCARTON; ALUISIO, 2010).

Cao et al. (2011) utilizam somente o conteúdo disponível de forma pública no *site* Download.com[4], sem a necessidade de informações pessoais dos autores das opiniões, da interação em redes sociais entre usuários e outras mais detalhadas, tais como, análise de subjetividade e polaridade, para concluir que as características semânticas apresentam maior influência.

Korfiatis et al. (2012) afirmam que testes de inteligibilidade são utilizados para quantificar diferentes tipos de texto dentro das áreas da ciência da informação e que as características textuais possuem maior influência na avaliação de utilidade das opiniões.

Além de chegarem a conclusões distintas utilizando técnicas semelhantes de processamento de texto, os autores utilizaram metodologias diferentes em dois pontos:

- Cao et al. (2011) utiliza a análise semântica latente (*LSA*, do inglês *Latent Semantic Analysis*) como ferramenta para criar um dos modelos de seu experimento;
- Korfiatis et al. (2012) utilizou métricas de inteligibilidade para verificar se o grau de dificuldade da compreensão da opinião pode influenciar no número de votos da opinião.

Essas abordagens diferentes criam uma lacuna para a investigação efetiva de quais características do texto são consideradas pelos usuários quando leem e avaliam uma

[1]Captura automática da nota sobre o produto e/ou serviço, data da publicação, etc.
[2]Cálculo do tamanho das palavras, tamanho das sentenças, quantidade de parágrafos, etc.
[3]Análise de significado das palavras do texto com métodos estatísticos.
[4]Plataforma *online* de divulgação de aplicativos para sistemas operacionais e celulares.

opinião como útil ou não. Portanto, a motivação desta pesquisa é tentar encontrar quais fatores influenciam no número de votos recebidos por uma opinião escrita em português.

1.2 OBJETIVOS

1.2.1 OBJETIVO GERAL

A presente dissertação adapta a metodologia descrita no trabalho de Cao et al. (2011) para o português, utiliza o domínio de serviços e estende a mesma pela utilização de métricas de inteligibilidade adaptadas à língua portuguesa (SCARTON; ALUISIO, 2010).

O domínio de *software* é originalmente utilizado no trabalho de Cao et al. (2011) e possui características de itens de experiência, ou seja, o usuário necessita utilizar antes de dar sua opinião. Porém, o domínio de serviços de hotéis foi escolhido pois se diferencia[5] do domínio original mas pertence a área de experiência de uso e é pouco explorado entre os trabalhos encontrados na literatura e utilizados como base desta dissertação.

O objetivo geral é identificar, dentre as características básicas, textuais e semânticas do conteúdo de opiniões de serviços de hotel escritas em português, quais mais influenciam as avaliações de utilidade dos usuários. Com a identificação das características mais influentes, é possível definir um modelo capaz de selecionar um conjunto de opiniões que possuem maior probabilidade de receber votos.

1.2.2 OBJETIVOS ESPECÍFICOS

1. Coletar os dados das opiniões sobre serviços de hotéis. Os textos das opiniões serão utilizados no cálculo das métricas de inteligibilidade básicas do Coh-Metrix-Port (SCARTON; ALUISIO, 2010). A segunda etapa será pré-processar o texto aplicando as técnicas de remoção de *stop words*[6] e de *stemming*[7];

2. Aplicar o método de regressão logística ordinal utilizado por Cao et al. (2011) nos modelos com as variáveis das características básicas, textuais, semânticas e algumas

[5]Serviços de hotéis possuem características como localização, atendimento, qualidade de serviços e *software* possui características como sistema operacional, facilidade de uso, versão.

[6]Palavras que podem ser consideradas irrelevantes para o conjunto de resultados exibido um sistema de busca.

[7]Radicalização ou determinação do radical. O objetivo é reduzir a variação das palavras de uma mesma raíz vocabular.

combinações: (i) básicas, (ii) textuais, (iii) semânticas, (iv) básicas e textuais e (v) básicas, textuais e semânticas.

3. Comparar os modelos utilizando as mesmas métricas aplicadas por Cao et al. (2011) e identificar entre as variáveis, qual grupo possui maior influência na quantidade de votos de utilidade.

1.3 ESTRUTURA DA DISSERTAÇÃO

Este documento está estruturado da seguinte forma:

Capítulo 2: Apresenta uma revisão conceitual de técnicas computacionais ligadas à área de mineração de opiniões utilizados na realização da pesquisa;

Capítulo 3: É dedicado à revisão bibliográfica, destacando-se a apresentação e análise de trabalhos correlatos às estratégias de captura e extração de informações automáticas das opiniões;

Capítulo 4: São apresentados a metodologia de seleção das informações, o tratamento dos dados e a criação dos modelos que foram utilizados na realização dos experimentos deste trabalho;

Capítulo 5: Descreve o processamento das opiniões nos experimentos e as métricas utilizadas para a comparação dos modelos;

Capítulo 6: Apresenta a análise e comparações dos resultados encontrados;

Capítulo 7: Finalizando este documento, apresenta as considerações finais onde são discutidos os resultados, as possíveis contribuições deste trabalho, suas limitações e algumas perspectivas de continuidade da pesquisa em trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 INTRODUÇÃO

A presente pesquisa está inserida na mineração de opiniões, especificamente na análise da qualidade da opinião. Neste capítulo será apresentada a fundamentação necessária à compreensão dos conceitos da análise semântica latente e da regressão logística ordinal. Também são discutidos os métodos de seleção de modelos, seleção de variáveis e as métricas de inteligibilidade utilizados nesta pesquisa.

O capítulo está organizado da seguinte maneira: a seção 2.2 traz os conceitos da análise semântica latente; a seção 2.3 trata sobre a regressão logística ordinal; o critério de informação de Akaike é descrito na seção 2.4; o método automático de seleção de variáveis é apresentados na seção 2.5; por fim, a definição das métricas de inteligibilidade implementadas para esta pesquisa são apresentadas na seção 2.6.

2.2 ANÁLISE SEMÂNTICA LATENTE

A análise semântica latente é um método de aprendizado matemático capaz de extrair e representar a semelhança de significados de palavras ou termos analisando uma grande quantidade de textos. No entanto, não é um método tradicional de processamento de linguagem natural ou inteligência artificial, pois não utiliza dicionários, bases de conhecimento, associações semânticas, gramáticas, analisadores sintáticos ou técnicas semelhantes (LANDAUER; DUTNAIS, 1997; LANDAUER et al., 1998).

O primeiro passo da *LSA* é representar o texto como uma matriz em que as linhas representam palavras distintas do conteúdo e as colunas são documentos em que as palavras estão inseridas (termos $\times$ documentos). Em seguida, as células da matriz são ponderadas por uma função que calcula a relevância das palavras no texto (DEERWESTER et al., 1990; LANDAUER et al., 1998).

O grau de relacionamento de uma palavra com o texto, denominado peso, indica

a importância da palavra em relação a um texto. A função de cálculo do peso utilizada na LSA é a TF-IDF (do inglês *term frequency–inverse document frequency*, que significa frequência do termo–inverso da frequência nos documentos). A função TF-IDF é definida por Jones (1972, 2004) e por Salton e Buckley (1988) como:

$$w_{ij} = tf_{ij} \times log\frac{N}{df_i} \tag{1}$$

onde, w_{ij} é o peso do termo i em um documento j, N é o número de documentos da coleção (*corpus*), tf_{ij} é a frequência do termo i no documento j e df_i é a quantidade de documentos da coleção que contém o termo i. Os termos que possuem os índices mais altos para o TF-IDF, normalmente são os que melhor caracterizam o tópico de um documento.

A dimensionalidade, ou seja, o número dos parâmetros pelos quais uma palavra ou sentenças são descritos, representa o relacionamento de palavras do mesmo contexto. É a redução da dimensionalidade que captura as palavras com significado semelhante e ocorrência em partes similares do texto.

Para reduzir o número dos parâmetros em um espaço vetorial com a menor perda de informação possível, a *LSA* se baseia no método de Decomposição em Valores Singulares (SVD, do inglês *Singular Value Decomposition*), proposto por Golub e Kahan (apud DEERWESTER et al., 1990). O SVD é uma técnica de manipulação algébrica de matrizes que re-orienta e ordena as dimensões em um espaço vetorial.

Seja A uma matriz termos-documentos resultante do primeiro passo da LSA, esse método à decompõe em três outras:

$$A = U \cdot \Sigma \cdot V^T \tag{2}$$

onde, U é uma matriz ortogonal unitária (LIPSCHUTZ; LIPSON, 2009) cujas colunas são os autovetores de $A \cdot A^T$; V é outra matriz ortogonal unitária, onde suas colunas são os autovetores de $A^T \cdot A$; e Σ é a matriz diagonal que possui os valores singulares de A, ou seja, as raízes quadradas não-negativas dos autovalores de $A \cdot A^T$ e $A^T \cdot A$, organizados em ordem decrescente. A figura 1 apresenta os passos da redução da dimensionalidade da matriz através da decomposição pelo SVD.

A melhor aproximação em k dimensões que se pode obter para a matriz é resultado da multiplicação das k primeiras linhas e colunas de Σ, as k primeiras colunas de U e k

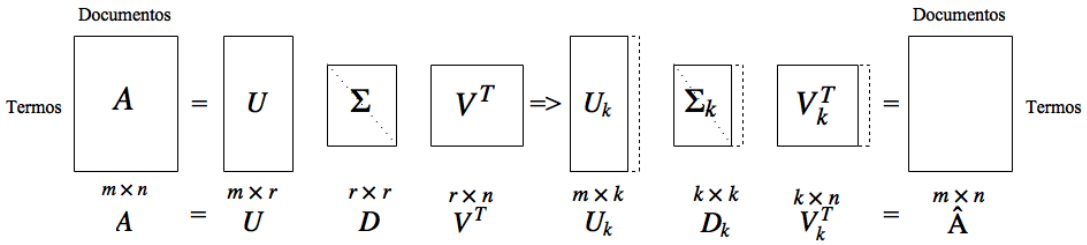

Figura 1: Redução da dimensionalidade através da SVD
Fonte: (DEERWESTER et al., 1990)

primeiras linhas de V. (DEERWESTER et al., 1990). Um vetor de documentos, isto é, o espaço semântico, contém uma representação LSA de cada documento.

Segundo Wiemer-Hastings et al. (2004), dentro da LSA é possível assumir que aproximadamente as primeiras 300 dimensões (de um universo de dezenas a centenas de milhares) são úteis para a captura do significado dos textos. Ao basear a representação em um número reduzido de dimensões, as palavras que ocorrem em contextos similares terão vetores com altos índices de similaridade. As dimensões descartadas podem ser consideradas ruído, associações aleatórias ou outros fatores não relevantes.

2.3 REGRESSÃO LOGÍSTICA ORDINAL

Um modelo de regressão é utilizado para avaliar o nível de relacionamento entre variáveis independentes e variáveis dependentes. Este modelo também é capaz de determinar a magnitude e a direção da influência das variáveis independentes sobre a variável dependente (CHEN; HUGHES, 2004).

Agresti (2002) apresenta duas categorias de modelos de regressão: os modelos de regressão linear e os modelos de regressão logística. A decisão de escolha entre um modelo ou outro depende da escala de medição da variável dependente, ou seja, se a variável dependente apresenta em sua escala um intervalo de valores então a regressão linear é a mais apropriada, caso a variável dependente seja binária/dicotômica, a regressão logística apresentará resultados mais significativos.

A regressão logística ordinal, do inglês *Ordinal Logistic Regression* ou OLR, é uma extensão do modelo de regressão logística, onde a variável dependente pode acomodar mais que duas categorias de valores. O modelo OLR é o mais apropriado para analisar o efeito de variáveis independentes em variáveis dependentes que possuem valores em uma escala ordinal, pois estas não podem assumir uma distribuição normal dos dados ou um intervalo

de valores.

Existem vários tipos de modelos de regressão logística ordinal que podem ser utilizados: (i) modelo de chances proporcionais, (ii) modelo de razão-contínua, (iii) modelo estereótipo e (iv) modelo de chances proporcionais parciais. Neste trabalho será apresentado e utilizado o modelo de chances proporcional (MCCULLAGH, 1980; ANDERSON, 1984; AGRESTI, 2012).

2.3.1 MODELO DE CHANCES PROPORCIONAL

O modelo de chances proporcional (MCP), do inglês *proportional odds model*, também chamado de modelo do logito cumulativo (*cumulative logit model*) é indicado quando a variável resposta era originalmente uma variável contínua que posteriormente foi agrupada (AGRESTI, 2002).

Esse modelo compara a probabilidade de uma resposta igual ou menor a uma determinada categoria ($j = 1, 2, ..., J-1$) com a probabilidade de uma resposta maior que esta categoria.

Na abordagem do modelo de chances proporcional são considerados $(J-1)$ pontos de corte das categorias sendo que o j-ésimo ($j = 1, ..., J-1$) ponto de corte é baseado na comparação de probabilidades acumuladas (AGRESTI, 2002, 2012):

$$\begin{aligned} logit[P(Y \leq j|x)] &= \log \frac{P(Y \leq j|x)}{P(Y > j|x)} = \log \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} \\ &= \log \frac{\pi_1(x) + \pi_2(x) + \cdots + \pi_j(x)}{1 - (\pi_{j+1}(x) + \pi_{j+2}(x) + \cdots + \pi_j(x))}, j = 1, 2, ..., J-1 \end{aligned} \tag{3}$$

Considerando-se as p covariáveis, a forma do modelo MCP é:

$$logit[P(Y \leq j|x)] = \alpha_j + \beta' x, j = 1, 2, ..., J-1 \tag{4}$$

O termo α é o intercepto do modelo e varia para cada uma das equações satisfazendo a condição $\alpha_1 \leq \alpha_2 \leq J \leq \alpha_{j-1}$; existem ainda p coeficientes betas (β) cujos elementos correspondem aos efeitos das covariáveis na variável resposta. É possível notar que β não depende de j, implicando que a relação entre x e Y é independente da categoria. Esse modelo fornece a estimativa de razão de chances (do inglês, *odds ratio*) para todas as categorias comparadas e que pode ser obtida exponenciando o coeficiente β.

Este modelo apresenta o mesmo efeito β para cada logito. Para um previsor x, a Figura 2 exemplifica o modelo quando $J = 4$.

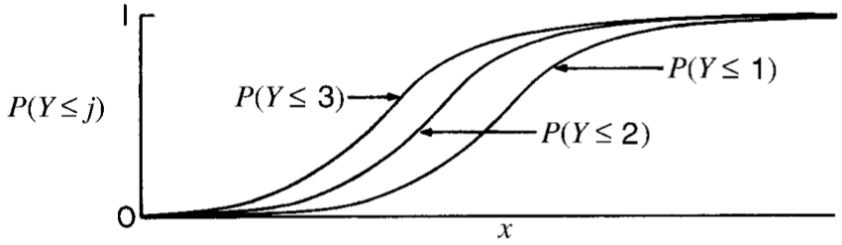

Figura 2: MCP com efeito independente do ponto de corte.
Fonte: (AGRESTI, 2002)

Para um determinado j, a curva de resposta é uma regressão logística binária com resultados $Y \leq j$ e $Y > j$. As curvas de resposta de $j = 1{,}2$ e 3 possuem o mesmo formato, pois compartilham a mesma taxa de crescimento e regressão, porém são horizontalmente distantes entre si. Para $j < k$, a curva para $P(Y \leq k)$ é a mesma que $P(Y \leq j)$ traduzida por $(\alpha_k$ - $\alpha_j)$ / β unidades na direção x (AGRESTI, 2002); ou seja,

$$P(Y \leq k|X = x) = P(Y \leq j|X = x + (\alpha_k - \alpha_j)/\beta). \tag{5}$$

As curvas das probabilidades das categorias são apresentadas na Figura 3. O MCP satisfaz a Equação 6.

$$\begin{aligned} &logit[P(Y \leq j|x_1)] - logit[P(Y \leq j|x_2)] \\ &= \log \frac{P(Y \leq j|x_1)/P(Y > j|x_1)}{P(Y \leq j|x_2)/P(Y > j|x_2)} = \beta'(x_1 - x_2) \end{aligned} \tag{6}$$

As chances de se obter uma resposta $\leq j$ em $x = x_1$ são $exp[\beta'(x_1 - x_2)]$ sobre as chances de $x = x_2$. O logito cumulativo para a razão de chances é proporcional a distância entre x_1 e x_2. A mesma proporcionalidade se aplica para cada um dos logitos do modelo.

2.3.2 SELEÇÃO DE VARIÁVEIS E AVALIAÇÃO DO MODELO

O ajuste do modelo OLR depende do número de variáveis independentes e da função de ligação. A função selecionada para o modelo descreve o efeito das variáveis independentes na variável ordinal dependente. Segundo Chen e Hughes (2004), o processo

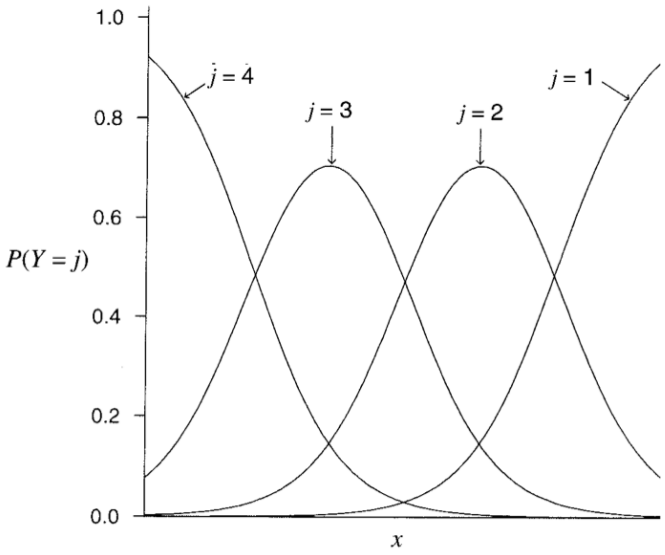

Figura 3: Propabilidade das categorias no MCP.
Fonte: (AGRESTI, 2002)

de seleção de variáveis ajuda a reduzir o risco de sobre ajuste do modelo reduzindo o número de variáveis independentes.

O principio da facilidade de interpretação e da parcimônia do modelo pode ser utilizado para encontrar modelos com ganhos mais significativos. O significado da parcimônia é a capacidade do modelo de não incluir algumas variáveis, isto é, excluindo-se algumas de suas variáveis independentes, as remanescentes são capazes de explicar seu resultado (CHEN; HUGHES, 2004).

Para avaliar a qualidade do ajuste dos modelos de regressão logística ordinal, normalmente são utilizados os testes de Pearson ou o desvio padrão. O teste de Pearson e a estatística de desvio padrão avaliam a adequação do ajuste comparando contagens observadas e esperadas (CHEN; HUGHES, 2004).

Estatísticas de ajuste como qui-quadrado de tendência e o critério de informação de Akaike também podem ser utilizados como medida de ajuste durante o processo de seleção de variáveis do modelo. Ambos os índices indicam que o modelo que possui os menores índices são os melhores modelos (CHEN; HUGHES, 2004; AGRESTI, 2012).

2.4 CRITÉRIO DE INFORMAÇÃO DE AKAIKE

Proposto por Akaike (1974), o critério de informação de Akaike (do inglês, Akaike Information criterion) ou AIC, é utilizado para manter o compromisso entre a precisão

do modelo e sua complexidade. Este método avalia um modelo sobre o quão próximo seu valores ajustados se aproximam dos valores reais em relação aos valores esperados.

Dada uma amostra, Akaike mostrou que seu critério é capaz de selecionar o modelo que maximiza:

$$AIC = -2(\text{máxima verossimilhança estimada} - \text{número de parâmetros do modelo}) \quad (7)$$

Seu cálculo penaliza a inclusão de muitas variáveis no modelo enquanto procura manter uma precisão razoável de maneira que, quanto menor o índice AIC, melhor o modelo.

2.5 SELEÇÃO AUTOMÁTICA DE VARIÁVEIS

Segundo Hocking (1976), vários métodos foram propostos para analisar pequenos subgrupos incluindo e removendo variáveis de acordo com critérios específicos. Estes procedimentos, normalmente conhecidos como métodos passo a passo (*no inglês, Stepwise*), consistem na variação de dois fundamentos básicos (FIELD, 2009): (i) a seleção para frente (no inglês, *Forward Selection*) e a seleção para trás (no inglês, *Backward Elimination*). Conforme Hocking (1976) e Field (2009) explicam:

Seleção para frente: A técnica de seleção para frente começa sem quaisquer variáveis no modelo de regressão. Para cada uma das variáveis preditivas candidatas, o método seleciona o previsor que apresenta o coeficiente de correlação mais alto e que reflete a contribuição que a variável traria para o modelo se fosse utilizada. Se esse previsor aumentar significativamente a capacidade do modelo prever a saída, ele é mantido (do contrário é descartado) e então é iniciada a busca pelo segundo previsor. Esse processo se repete até que todas as variáveis candidatas tenham sido avaliadas.

Seleção para trás: Essa técnica é o oposto do método de seleção para frente, já que começa em um modelo com todas as variáveis preditivas. A partir de então, calcula-se a contribuição de cada previsor, verificando-se a significância do teste t de cada variável. O valor da significância do previsor é comparado com um critério de remoção e caso satisfaça esse critério, é removido do modelo e o calculo é novamente estimado para os previsores restantes.

Segundo Field (2009), o método de seleção avançada apresenta menor probabilidade de inclusão de variáveis que podem causar interferência na capacidade preditiva do modelo, porém ainda assim existe a possibilidade deste método eliminar um previsor que de fato teria uma melhor contribuição para o modelo.

2.6 MÉTRICAS DE INTELIGIBILIDADE

O termo inteligibilidade está relacionado com as características do texto que são determinantes para que a leitura seja facilitada, isto é, o uso de palavras frequentes e estruturas sintáticas simples. É construída pelo leitor através da sua interação com o texto e seu autor, buscando a compreensão do conteúdo através do processamento linguístico: léxico, relações sintáticas, sinais de pontuação (SCARTON; ALUISIO, 2010).

A compreensão pode ser facilitada se as relações entre as partes do texto estiverem adequadas, ou seja, utilizam conjunções, advérbios e elementos de correferência (como pronomes, elipses, entre outros). A legibilidade está relacionada aos aspectos gráficos que contribuem na construção do sentido, como as ilustrações, os diagramas, as fotografias, o formato e cor das letras.

Existem somente duas ferramentas de avaliação da inteligibilidade para a língua portuguesa (SCARTON; ALUISIO, 2010): (i) a fórmula adaptada para o português (MARTINS et al., 1996) do índice *Flesh Reading Ease* e (ii) a adaptação para a língua portuguesa das métricas do Coh-Metrix (MCNAMARA et al., 2002; GRAESSER et al., 2004; CROSSLEY et al., 2007) proposto por pesquisadores da USP/São Carlos em Scarton et al. (2010).

2.6.1 ÍNDICE DE INTELIGIBILIDADE

O índice *Flesch Reading Ease* (MARTINS et al., 1996) avalia superficialmente a inteligibilidade de um texto. Esta métrica mede somente o número de palavras em sentenças e o número de letras ou sílabas por palavra. A saída da fórmula *Flesch Reading Ease* é um número entre *0* e *100*, e quanto mais alto o índice, mais fácil a leitura (BARBOZA; NUNES, 2007; SCARTON; ALUISIO, 2010):

$$206.835 - (1.015 \times ASL) - (84.6 \times ASW) \tag{8}$$

onde, *ASL* é o tamanho médio de sentenças medido em número de palavras (o

número de palavras dividido pelo número de sentenças) e ASW é número médio de sílabas por palavra (o número de sílabas dividido pelo número de palavras). Em português, a adaptação do *Flesch Reading Ease* resultou na fórmula:

$$248.835 - (1.015 \times ASL) - (84.6 \times ASW) \tag{9}$$

que corresponde a fórmula original salvo pela constante 248.835 que diferencia textos em inglês de textos em português[1]. De acordo com (SCARTON; ALUISIO, 2010), os valores desse índice variam entre de 0-100 conforme a tabela 1:

Tabela 1: Os graus de escolaridade e a facilidade de compreensão dos textos

Facilidade de compreensão	**Flesch Reading Ease**	**Grau de escolaridade**
Muito difícil	0 - 30	Universitário
Difícil	30 - 50	Nível Médio ou universitário incompleto
Mais difícil	50 - 60	Nível Médio incompleto
Padrão	60 - 70	7a ou 8a Série
Mais fácil	70 - 80	6a Série
Fácil	80 - 90	5a série
Muito fácil	90 - 100	4a série

Fonte: (SCARTON; ALUISIO, 2010)

2.6.2 A FERRAMENTA COH-METRIX-PORT

A ferramenta Coh-Metrix é utilizada para calcular os índices de coesão, coerência e a dificuldade de compreensão de um texto (em inglês) por meio de análise linguística de diferentes tipos: léxico, sintático, discursivo e conceitual. A coesão é definida por Graesser et al. (apud SCARTON; ALUISIO, 2010) como "*características de um texto que, de alguma forma, ajudam o leitor a conectar mentalmente as ideias do texto*".

A coerência também é definida por Graesser et al. (apud SCARTON; ALUISIO, 2010) como "*características do texto (ou seja, aspectos de coesão) que provavelmente contribuem para a coerência da representação mental*". Para todas as métricas, vários recursos de PLN são utilizados. Os 60 índices estão divididos em seis classes principais que são (SCARTON; ALUISIO, 2010):

[1]Segundo Martins et al. (1996), 42 é o valor de referência que, na média, diferencia textos em inglês e português.

- Identificação Geral e Informação de Referência: corresponde às informações que referenciam o texto, como título, gênero entre outros.
- Índices de Inteligibilidade: contém os índices de inteligibilidade calculados com as fórmulas Flesch Reading Ease e Flesch Kincaid Grade Level.
- Palavras Gerais e Informação do Texto: é dividida em subclasses, tais como, contagens básicas, frequências, concretude e hiperônimos.
- Índices Sintáticos: é subdividida nas classes: constituintes, pronomes, tipos e tokens, conectivos, operadores lógicos e similaridade sintática de sentenças.
- Índices Referenciais: esta classe possui três subclasses: anáfora, correferência e LSA.
- Semânticos e Dimensões do Modelo de Situações: classe que possui as subclasses: dimensão causal, dimensão intencional, dimensão temporal e dimensão espacial.

A versão adaptada para o português, o Coh-Metrix-Port, além de utilizar o índice Flesch Reading Ease (MARTINS et al., 1996), também possui as seguintes 40 métricas:

- Contagens Básicas (13): número de palavras, número de sentenças, número de parágrafos, sentenças por parágrafos, palavras por sentenças, sílabas por palavras, número de verbos, número de substantivos, número de advérbios, número de adjetivos, número de pronomes, incidência de palavras de conteúdo (substantivos, adjetivos, advérbios e verbos) e incidência de palavras funcionais (artigos, preposições, pronomes, conjunções e interjeições).
- Constituintes (3): incidência de sintagmas nominais, modificadores por sintagmas nominais e palavras antes de verbos principais.
- Frequências (2): frequência de palavras de conteúdo e mínimo das frequências de palavras de conteúdo.
- Conectivos (9): incidência de todos os conectivos, incidência de conectivos aditivos positivos, incidência de conectivos temporais positivos, incidência de conectivos causais positivos, incidência de conectivos lógicos positivos, incidência de conectivos aditivos negativos, incidência de conectivos causais negativos, incidência de conectivos temporais negativos e incidência de conectivos lógicos negativos.
- Operadores Lógicos (5): incidência de operadores lógicos, número de *e*, número de *ou*, número de *se* e número de negações.

- Pronomes, Tipos e Tokens (3): incidência de pronomes pessoais, pronomes por sintagmas e relação tipo/token.
- Hiperônimos (1): hiperônimos de verbos.
- Ambiguidades (4): ambiguidade de verbos, de substantivos, de adjetivos e de advérbios

2.7 CONCLUSÃO

Este capítulo apresentou os principais conceitos utilizados para a construção desta pesquisa. Sendo destacados a análise semântica latente, a regressão logística ordinal, os métodos de automáticos de seleção de variáveis, o método de seleção de modelos e as métricas de inteligibilidade disponíveis na língua portuguesa que serão utilizadas nos experimentos conduzidos por esta pesquisa.

O próximo capítulo trará uma visão geral sobre a análise de sentimentos e a qualidade da opinião encontrados na literatura, realizando uma comparação entre eles.

3 MINERAÇÃO DE OPINIÕES

3.1 INTRODUÇÃO

A mineração de opiniões é uma área da mineração de dados que utiliza técnicas computacionais para extrair, classificar, entender e acessar opiniões de fontes como fóruns, *blogs*, comentários em redes sociais e outros conteúdos gerados por usuários.

Este capítulo apresenta um estudo sobre a mineração de opiniões, define o problema da mineração de opiniões e as etapas da e abordagens utilizadas para solucioná-lo. Por fim, são apresentados alguns trabalhos encontrados na literatura e que relacionam-se com os objetivos desta pesquisa.

O restante do capítulo está organizado da seguinte maneira: a seção 3.2 aborda a definição da mineração de opiniões de uma maneira geral; e a seção 3.3 apresenta a revisão bibliográfica da literatura sobre a qualidade da opinião.

3.2 MINERAÇÃO DE OPINIÕES

A análise de sentimentos ou mineração de opiniões é o estudo de opiniões, emoções e sentimentos expressos em textos utilizando técnicas computacionais e realizada sobre textos de quaisquer tamanho e formato, tais como páginas *web*, *posts*, *tweets* e comentários sobre produto e/ou serviços (LIU, 2010).

Para Liu (2010), as opiniões são expressões subjetivas que descrevem avaliações ou atitudes pessoais sobre entidades, eventos e suas propriedades. Zhang e Varadarajan (2006) definem uma boa opinião como uma combinação equilibrada entre uma avaliação subjetiva e informações objetivas. Opiniões totalmente neutras ou indiferentes são inúteis; assim como opiniões bem escritas podem ser úteis e convincentes.

3.2.1 DEFINIÇÃO DO PROBLEMA

Quando uma pessoa precisa decidir sobre algo, normalmente procura a opinião de um amigo ou um parente. O mesmo é válido para empresas que consultam seus consumidores sobre a qualidade de seus produtos. A constante interação e a facilidade de criação de conteúdo *online* contribuiu na mudança da forma como os usuários expressam seus pontos de vista e opiniões. As pessoas passaram a publicar suas opiniões diretamente no *site* do comerciante ou relatar suas experiências em fóruns de discussão, *blogs* e redes sociais (CHEN; ZIMBRA, 2010; LIU, 2010).

Esse interação virtual representa uma fonte de informação gerada por usuários que tiveram contato com produtos ou serviços, o que para uma pessoa comum significa alcançar avaliações que vão além de parentes ou amigos. Da mesma forma que para empresas, o conteúdo gerado por usuários, pode significar maior agilidade na obtenção de informações sobre suas ações de mercado (LIU, 2010).

No entanto, ainda segundo Liu (2010), esse monitoramento da opinião em diversas fontes de dados é uma tarefa complexa devido ao grande volume de conteúdo que pode ser recuperado. É possível que muitas opiniões estejam escondidas em *sites*, longas publicações de *blogs* e redes sociais, o que dificulta a descoberta de informações relevantes, a extração de textos com avaliações, a leitura e a organização em um formato mais simples de ser manipulado.

3.2.2 DEFINIÇÃO DA OPINIÃO

Uma opinião sobre um objeto é uma entidade associada a um produto, pessoa, evento, empresa ou assunto que é definida pela relação entre os itens que compõem o objeto e suas características. A expressão "objeto" é utilizada para definir a entidade que é alvo da opinião. As palavras "tópico" ou "aspecto" são usualmente relacionados aos objetos do domínio de eventos (ou serviços) e o termo "característica" é normalmente utilizado no domínio de opiniões sobre produtos (CHEN; ZIMBRA, 2010; LIU, 2010; TSYTSARAU; PALPANAS, 2010)

O autor ou a fonte da opinião é a pessoa ou organização que expressou a opinião em um determinado momento no tempo. Um documento (ex. avaliações de produtos, publicações em fóruns e *blogs*) é definido por uma sequência de sentenças e a opinião sobre uma característica de um objeto é definida por um grupo de sentenças consecutivas que possuem avaliações positivas ou negativas relacionadas à característica (LIU, 2012).

A orientação semântica de uma característica determina a polaridade da opinião, que pode ser positiva ou negativa (CHEN; ZIMBRA, 2010; LIU, 2010). A polaridade da opinião é encontrada através da classificação das palavras do texto em positivas e negativas. Textos com avaliações neutras normalmente são encontrados em notícias ou conteúdos apenas informativos (LIU, 2012).

Chen e Zimbra (2010) definem um objeto como uma representação finita de características, incluindo o próprio objeto como um elemento especial. Cada característica pertence a um conjunto finito de características pré-definidas do objeto que também pode ser representada por seus sinônimos (ex. resolução ou qualidade da imagem de uma câmera digital). Um documento com opiniões possui avaliações sobre um conjunto de objetos de um conjunto de autores.

3.2.3 OPINIÕES OBJETIVAS, SUBJETIVAS E COMPARATIVAS

Uma opinião objetiva possui informações que enumeram as características dos produtos e/ou serviços e servem como uma descrição alternativa para confirmar (ou rejeitar) as informações anunciadas pelas empresas. Uma opinião subjetiva descreve um item com impressões pessoais e pode incluir informações que não aparecem na descrição oficial do mesmo (GHOSE; IPEIROTIS, 2007; LIU, 2010). A Figura 4 é um exemplo de uma opinião subjetiva.

Uma opinião comparativa relaciona a semelhança ou a diferença entre dois ou mais objetos baseado na preferência do autor por determinada característica comum aos objetos comparados. Normalmente a avaliação é realizada através do modo comparativo ou superlativo de adjetivos e advérbios, porém, isto não é uma regra.

Segundo Ghose e Ipeirotis (2007), as opiniões objetivas são a preferência em relação à bens duráveis (ex. eletrônicos), pois de certo modo confirmam as informações da descrição e possuem uma interpretação mais simples ao descrever o modo de uso ou uma deficiência de um item.

Para artigos de experiência (ex. filmes, livros, serviços), os usuários tem preferência por uma breve descrição “objetiva” sobre algumas características e uma descrição pessoal mais detalhada e emotiva que consiga descrever aspectos que uma revisão objetiva não conseguiria apresentar. A polaridade aproximada (i.e. positiva ou negativa) da opinião pode ser extraída da nota já atribuída ao conteúdo pelos usuários (GHOSE; IPEIROTIS, 2007).

Figura 4: Exemplo de opinião subjetiva.
Fonte: Tripadvisor

Liu (2010) ressalta que ao estudar a subjetividade das emoções ou opiniões, é necessário considerar a diferença entre duas noções importantes: (i) o estado mental (sentimento) de uma pessoa e (ii) a linguagem utilizada para descrever este estado.

Segundo Ekman (1992), as emoções humanas são sentimentos e sensações subjetivas, que embora limitadas em 6 tipos básicos, ou seja, tristeza, medo, surpresa, repulsa, raiva e alegria, podem ser escritos de uma forma bastante ampla com expressões de linguagem. A análise de sentimentos ou mineração de opiniões, tenta inferir sobre os sentimentos das pessoas baseado somente na linguagem textual utilizada.

3.2.4 ETAPAS DA MINERAÇÃO DE OPINIÃO

A mineração de opiniões pode ser dividida nas seguintes tarefas: (a) identificar e processar as opiniões sobre determinados assuntos ou objetos em um conjunto de documentos; (b) classificar a orientação semântica ou polaridade das opinião; e (c) apresentar os resultados de forma simplificada.

3.2.4.1 IDENTIFICAÇÃO DAS OPINIÕES

Considerando um conjunto de textos extraídos de alguma fonte (e.g. jornais, redes sociais, agregadores de opinião produtos e/ou serviços), a etapa de identificação tem por objetivo encontrar os tópicos e aspectos existentes nas opiniões e criar uma associação com o conteúdo subjetivo.

Normalmente todo documento (desde sentenças até o conjunto de textos) refere-se a uma única entidade como alvo da opinião e dessa forma, o principal desafio está em identificar os aspectos desta entidade (PANG; LEE, 2008). Esta tarefa pode incluir também a separação entre conteúdo ou sentenças que possuem ou não opiniões, com o objetivo de melhorar o resultado da tarefa de classificação de polaridade.

3.2.4.2 CLASSIFICAÇÃO DE POLARIDADE

A classificação de polaridade também pode ser um problema de classificação binária, isto é, a classificação do texto em duas classes, ou seja, uma positiva e outra negativa. Porém, classes adicionais podem ser consideradas para que a análise seja mais completa, ou para aumentar o nível de detalhes dos resultados com diferentes graus de intensidade (ex. “excelente”, “razoável”) ou intervalos numéricos que representam uma escala de valores (ex. 0 à 5) (TSYTSARAU; PALPANAS, 2010).

Outra abordagem é considerar a categoria neutra, que engloba textos sem uma tendência clara quanto a sua polaridade ou simplesmente sem sentimento (notícias, textos científicos). Porém um texto neutro é diferente de um texto não polarizado. Em um texto não polarizado, não existe elementos suficientes para classificá-lo, e dessa forma a tarefa de classificação não consegue chegar à conclusão sobre sua polaridade. Isso acontece em conteúdos com erros tipográficos e sentenças incompletas (PANG; LEE, 2008; TSYTSARAU; PALPANAS, 2010).

3.2.4.3 SUMARIZAÇÃO

Para identificar a opinião média ou prevalecente de um grupo de pessoas sobre um determinado tópico/entidade, é necessário analisar uma grande quantidade de opiniões (LIU, 2012). Em um conjunto de opiniões um sumário de um determinado produto e/ou serviço pode ajudar um consumidor a identificar seus respectivos pontos fortes e fracos de maneira mais rápida e ainda levando em consideração a experiência prévia de outras pessoas (ex: Figura 5).

O sentimento sumarizado também pode ser utilizado de diversas formas, como prever eleições, comportamento da bolsa de valores, arrecadação de bilheterias de filmes, definição de preços, etc (CHEN; ZIMBRA, 2010; LIU, 2012).

Figura 5: Um resumo das opiniões sobre um hotel
Fonte: TripAdvisor

3.3 REVISÃO BIBLIOGRÁFICA

A revisão do estado da arte desta pesquisa foi realizada a partir de buscas nas base de publicações cientificas do *IEEE*, *ACM*, Cite Seer X, *Web Of Knowledge* e Periódicos Capes com palavras-chave relacionadas a mineração de opiniões, análise da qualidade, utilidade da opinião e qualidade da opinião (utilizando suas equivalentes em inglês[1]).

A partir dos resultados encontrados e da pesquisa sobre mineração de opiniões criado por Liu (2012), foram selecionados trabalhos relacionados a qualidade da opinião no período de 2006 à 2013 ordenados por número de citações, além de alguns trabalhos mais recentes. As pesquisas apresentadas nesta seção buscaram diversas formas de estimar a influência do conteúdo criado pelos usuários na avaliação utilidade das opinião publicadas *online* a respeito de produtos e/ou serviços.

A seleção dos artigos deixou de fora trabalhos relacionados a opiniões que utilizam o relacionamento de usuários em redes sociais, pois o objetivo foi utilizar somente as informações publicamente disponíveis em lojas virtuais, *sites* agregadores de conteúdo e catálogos de filmes e/ou serviços.

[1] *opinion mining*, *sentiment analysis*, *review quality*, *review helpfulness* e *opinion quality*.

3.3.1 CARACTERÍSTICAS BÁSICAS

Algumas pesquisas anteriores (KIM et al., 2006; ZHANG; VARADARAJAN, 2006; GHOSE; IPEIROTIS, 2007; LIU et al., 2007, 2008; O'Mahony; SMYTH, 2009; TSUR; RAPPOPORT, 2009; LU et al., 2010), se basearam na tentativa de determinar automaticamente a qualidade (ou a utilidade) das opiniões utilizando características textuais. O problema de determinar a qualidade da opinião é formulado como um problema de classificação ou regressão, onde o voto dos usuários é utilizado como o valor de referência.

Neste contexto, Zhang e Varadarajan (2006), coletaram as opiniões e os votos de utilidade atribuídos por usuários em produtos da loja *online* Amazon.com. Baseado nesses votos, os autores atribuíram a qualidade da opinião a razão entre o número de pessoas que votaram sobre o número de pessoas que leram a opinião. Para classificar a qualidade das opiniões, os autores utilizaram o modelo de regressão SVM (máquina de vetores de suporte, ou do inglês, *support vector machine*). A função de regressão resultante foi utilizada para classificar a qualidade de novas opiniões. Os resultados alcançados mostraram que as características textuais superficiais (i.e. contagem de substantivos, verbos, adjetivos) podem ser úteis na estimativa da qualidade das opiniões.

Kim et al. (2006) também empregou um modelo de regressão SVM para estimar a utilidade das opiniões. Sua função de regressão foi construída a partir de características estruturais (ex. comprimento do texto, sentenças e marcadores *HTML*), características léxicas (ex. uni-gramas, bi-gramas, radicais das palavras), textuais, semânticas (capturadas com auxílio de dicionários) e meta-dados (ex. nota/número de estrelas em avaliações) das opiniões analisadas. Os autores perceberam que as características léxicas, o tamanho do texto e a nota atribuída ao produto tem influência da avaliação da qualidade das opiniões.

3.3.2 CARACTERÍSTICAS TEXTUAIS

Utilizando opiniões do *site* TripAdvisor, Talwar et al. (2007) também considerou as características textuais na tentativa de identificar sobre quais aspectos dos serviços os usuários estavam discutindo. Esta pesquisa explorou a relação entre uma opinião e suas antecessoras. Segundo (TALWAR et al., 2007), a leitura das opiniões *online* mostra que as notas atribuídas por outros usuários são muitas vezes parte de tópicos de discussão, isto é, uma publicação não é, necessariamente, independente da outra.

Liu et al. (2007) criou e utilizou um conjunto de treinamento classificado manualmente para identificar opiniões com pouca qualidade no domínio de produtos eletrônicos. Seu objetivo foi aperfeiçoar um método automático para resumir as avaliações e relacioná-las as propriedades dos produtos. Ghose e Ipeirotis (2007) combinaram um modelo de vendas com a análise textual e demostraram que opiniões extremas (muito positivas ou negativas em relação a nota atribuída ao alvo da opinião) são consideradas mais úteis.

Em Liu et al. (2008), os autores utilizaram a experiência do autor da opinião, seu histórico de avaliações e o estilo de escrita em um modelo de regressão não linear para prever a utilidade das opiniões. Sua análise se baseou em opiniões de filmes extraídos do IMDB[2] (Internet Movie Database), porém argumentam que sua abordagem suficientemente genérica e capaz de ser aplicada a qualquer domínio. No entanto, capturar o estilo de escrita e a experiência do autor da opinião requer acesso a informações restritas, o que dificulta a utilização desse método.

O'Mahony e Smyth (2009) treinaram um classificador de utilidade para opiniões de serviços de hotel a partir de um grande conjunto de opiniões com o intuito de classificar todo tipo de opinião, incluindo as que nunca receberam avaliações prévias sobre sua utilidade. Na pesquisa foram utilizadas quatro categorias de características das opiniões: (i) reputação (conjunto de opiniões que os usuários criaram anteriormente), (ii) conteúdo (número de termos no texto da opinião), (iii) social (extraído da distribuição de opiniões usuário-hotel) e (iv) sentimento (quanto os usuários aproveitaram a experiência com o hotel). Os autores chegaram a conclusão que a reputação e o sentimento foram as características mais importantes para a classificação, inclusive, a performance do resultado se manteve alto, mesmo sem a utilização da reputação, o que pode acontecer em casos onde estas informações não estão disponíveis.

Tsur e Rappoport (2009) apresentaram o algoritmo *RevRank*, criado para organizar as opiniões automaticamente se baseando no número de votos de utilidade das opiniões. Seu método não-supervisionado procurou eliminar o trabalho e a possibilidade de ocorrência de falhas na classificação manual de bases de treinamento. A identificação da frequência e dos termos mais importantes foi realizada com técnicas de PLN e um *corpus* com palavras comuns na língua inglesa.

Lu et al. (2010) investigou como o contexto social das opiniões pode melhorar a eficiência da previsão de utilidade através de aprendizado não-supervisionado, contando com uma pequena base de treinamento já classificada e uma grande quantidade de dados

[2]Catalogo online de avaliação de filmes

sem classificação sobre celulares, artigos de beleza e câmeras digitais.

3.3.3 CARACTERÍSTICAS SEMÂNTICAS

Segundo (LU et al., 2010), sua pesquisa foi uma das primeiras a combinar informações textuais e a rede social do autor da opinião para definir a utilidade do conteúdo do texto. A fonte de dados foi *site* Britânico Ciao UK[3] e que utilizou as características: (i) textuais (tamanho da opinião, média do tamanho das sentenças, variedade do vocabulário), (ii) sintáticas (porcentagem de pronomes, adjetivos e pontuação), (iii) conformidade (medida de quanto a opinião é semelhante a média das opiniões), (iv) sentimento (palavras positivas ou negativas da opinião) e (v) contexto social (engajamento dos autores das opiniões, histórico da qualidade, *status* do autor na rede social). Estas características, porém são semelhantes as exploradas por (O'Mahony; SMYTH, 2009), e não foi possível perceber avanços nos métodos, apesar dos autores da pesquisa afirmarem que os algoritmos podem ser utilizados em outros domínios e que são efetivos mesmo quando o contexto social não está disponível.

No ano de 2011, Ghose e Ipeirotis (2011) retornam ao problema da utilidade das opiniões se baseando em sua própria pesquisa anterior (GHOSE; IPEIROTIS, 2007). Porém, dessa vez, foram realizados múltiplos níveis de análise de texto para identificar quais características das opiniões são importantes. O estudo foi realizado sobre características lexicais, gramaticais, semânticas, estilo de escrita e o histórico dos autores das opiniões para identificar quais destas possuem maior influência na percepção de utilidade e índice de vendas.

Ao invés de prever o nível de utilidade de opiniões que não possuem votos, Cao et al. (2011) investigou os fatores que determinaram a quantidade de votos de utilidade que as opiniões receberam (incluindo tanto votos "sim" quanto "não"), examinando os efeitos de características básicas, o estilos de escrita e a semântica das palavras. Foram utilizadas mais de 3.400 opiniões sobre *software* de computador (entre gratuitos e/ou pagos) publicadas no *site* Download.com.

Segundo Cao et al. (2011), as características básicas são informações que podem ser observadas diretamente, como a nota sobre produtos ou serviço associadas a opinião e a data da publicação. O estilo de escrita representa algumas características chave na redação do texto, como sentenças curtas com palavras simples ou sentenças longa com palavras mais sofisticadas. Por fim, as características semânticas (extraídas através da

[3]http://www.ciao.co.uk

LSA) referem-se ao significado que as palavras possuem na opinião, ou seja, são as palavras que possuem maior probabilidade de influenciar o voto dos leitores.

A conclusão do trabalho de Cao et al. (2011), sugere que avaliações com *opiniões extremas* (i.e. opiniões muito diferentes da média geral) e as *características semânticas* das opiniões tem mais impacto do que as outras características no número de votos que as opiniões recebem.

Korfiatis et al. (2012) explorou as características textuais de uma base de dados contendo 36.856 opiniões sobre livros da loja Amazon UK[4], empregando métricas de inteligibilidade e concluíram que a utilidade da opinião é influenciada pelo estilo da escrita. Korfiatis et al. (2012) afirma que pessoas que enviaram suas opiniões tiveram experiência de consumo com o produto, por esse motivo poderiam contribuir com sua opinião pessoal (negativa ou positiva, dependendo da nota atribuída pelo usuário). Os pesquisadores consideram que a opinião é utilizada como "justificativa" para o valor da nota do produto, deixando o potencial comprador decidir se a opinião foi útil ou não.

Para acrescentar valor as opiniões online e resolver os problemas relacionados com a qualidade e a credibilidade, alguns *sites* que capturam opiniões sobre seus produtos ou serviços permitem que os usuários "avaliem as avaliações". A principal abordagem é perguntar se a opinião é útil ou não (LI et al., 2013). Um índice de utilidade pode ser calculado utilizando a porcentagem de votos úteis sobre todos os votos. Esse tipo de atribuição de valor representa um tipo de certificado de qualidade e permite ao leitor encontrar rapidamente o conteúdo mais útil em meio a centenas de opiniões.

Segundo Chen et al. (2008) e Chen e Huang (2013), opiniões com um grande número de votos de utilidade possuem correlação positiva com o número de vendas. Além de ser um ítem relacionado a qualidade, o voto de utilidade também pode ser associado a uma atribuição subjetiva de valor de uma informação presente na opinião (CAO et al., 2011).

3.4 CONCLUSÃO

Este capítulo apresentou algumas considerações sobre a qualidade e a percepção de utilidade das opiniões. O estudo buscou, em linhas gerais, determinar as principais diferenças existentes entre as características que influenciam a percepção de utilidade das opiniões, e nesse cenário posicionou as técnicas utilizadas na presente pesquisa.

[4]http://www.amazon.co.uk

Foi possível perceber que as técnicas utilizadas evoluíram ao longo do tempo e recentemente, a maioria das pesquisas procura as características textuais e semânticas das opiniões. O próximo capítulo apresenta a metodologia adotada na pesquisa, faz sua modelagem e detalha os elementos necessários aos experimentos realizados.

4 MÉTODO PROPOSTO

4.1 INTRODUÇÃO

Nos capítulos anteriores foram apresentados a fundamentação teórica necessária à compreensão da mineração de opiniões e uma visão geral sobre o problema da análise da qualidade das opiniões, incluindo as diversas abordagens encontradas na literatura. Neste capítulo será apresentado a descrição do método proposto para a realização do experimento de avaliação das características que influenciam nos votos de utilidade das opiniões derivado da metodologia utilizada por Cao et al. (2011).

O restante do capítulo está organizado da seguinte maneira: a seção 4.2 descreve os dados coletados; a seção 4.3 apresenta as características da opinião do *site* TripAdvisor; as seções 4.4, 4.5 e 4.6 detalham respectivamente as características básicas, textuais e semânticas que foram utilizadas no experimento; e finalmente a seção 4.7 apresenta os 5 modelos avaliados no experimento.

4.2 COLETA DE DADOS

A captura automática das opiniões do TripAdvisor resultou numa amostra contendo 35.037 opiniões de 2.118 hotéis das 12 cidades do Brasil que foram selecionadas para sediarem os jogos da Copa do Mundo de 2014: Rio de Janeiro (RJ), São Paulo (SP), Belo Horizonte (MG), Porto Alegre (RS), Brasília (DF), Cuiabá (MT), Curitiba (PR), Fortaleza (CE), Manaus (AM), Natal (RN), Recife (PE) e Salvador (BA).

Apesar de domínios diferentes, em Cao et al. (2011) foram utilizadas 3.460 opiniões e em Korfiatis et al. (2012) pouco mais de 36 mil opiniões. A quantidade de opiniões (absoluta e porcentual) que os hotéis de cada cidade da copa receberam é apresentada na tabela 2.

Conforme discutido no trabalho de Cao et al. (2011), a maioria das opiniões não recebe voto de utilidade. O problema que se repete em outros trabalhos citados

Tabela 2: Quantidade de opiniões por cidade.

Cidade	Hotéis (Qtd.)	Opiniões (Qtd.)	%
Belo Horizonte	138	2.142	6,11
Brasília	79	2.230	6,36
Cuiabá	50	545	1,56
Curitiba	130	2.873	8,2
Fortaleza	184	2.853	8,14
Manaus	66	841	2,4
Natal	188	2.778	7,93
Porto Alegre	98	1.823	5,2
Recife	90	1.837	5,24
Rio de Janeiro	452	6132	17,5
Salvador	199	2.684	7,66
São Paulo	444	8.299	23,69
Total	2.118	35.037	100

Fonte: Autoria própria.

anteriormente, também foi encontrado na distribuição de votos por opinião neste trabalho (Tabela 3).

Tabela 3: Quantidade de opiniões por votos de utilidade.

Votos de Utilidade	Opiniões (Qdt.)	%
0	21.686	61,89
1	8.186	23,36
2	2.825	8,06
3	1.132	3,23
4	508	1,45
5	271	0,77
6	147	0,42
≥ 7	282	0,8
Total	35.037	100

Fonte: Autoria própria.

É possível notar que o aumento do número de votos implica em uma diminuição expressiva da quantidade de opiniões, confirmado através do índice de Correlação de Pearson ($-0,77$), que foi encontrado utilizando-se a quantidade de votos e opiniões como variáveis de entrada. A quantidade de opiniões que nunca recebeu voto ultrapassa 60% da base, enquanto a quantidade que recebeu sete ou mais votos, não alcança 1% do total de opiniões.

Também foi observado, conforme apresentado na Tabela 4, que mais da metade dos hotéis recebe no máximo 10 opiniões.

Tabela 4: Número de hotéis agrupados por número de opiniões.

Número de Opiniões (N)	Número de Hotéis	%
$N \leq 10$	1313	61.99
$10 < N \leq 30$	448	21.15
$30 < N \leq 100$	311	14.68
$N > 100$	46	2.17
Total	2118	100

Fonte: Autoria própria.

4.3 OPINIÃO DOS USUÁRIOS DO TRIPADVISOR

A Figura 6 é um exemplo de uma opinião registrada no TripAdvisor. No canto superior esquerdo abaixo do título está a avaliação geral do hotel (indicada pelo número 1 em destaque) realizada pelo usuário, seguida pela data de publicação da opinião.

Na sequência é apresentada o texto livre da opinião. A data da hospedagem no hotel e as avaliações das características do hotel (destaque número 2), quando o usuário preenche esses dados, estão posicionados antes da pergunta sobre a utilidade da opinião (circulado no número 3). O TripAdvisor apresenta somente o número total de votos que a opinião recebeu, não havendo diferença entre votos positivos ou negativos.

4.4 CARACTERÍSTICAS BÁSICAS DAS OPINIÕES

As características básicas, encontradas anteriormente no exemplo da Figura 6, podem ser extraídas de forma simples e sem a utilização de ferramentas de PLN. Foram utilizados: o intervalo (em dias) entre a data da publicação da opinião e a data da viagem; e o "extremismo" da avaliação definido como o valor estimado pela diferença entre a nota da opinião dada ao hotel e média geral de todos os usuários.

Além desses itens, também foram utilizadas as notas em uma escala de 1 a 5 para as características: custo-benefício, localização, qualidade do sono, quartos, limpeza e atendimento. A escala das notas das características foi convertida em um valor binário, onde a avaliação é positiva para as notas $> 2,5$ e negativa para notas $\leq 2,5$.

4.5 CARACTERÍSTICAS TEXTUAIS DAS OPINIÕES

As características textuais representam o estilo de escrita do usuário e não podem ser extraídas sem um pré-processamento no conteúdo da opinião. Para esta dis-

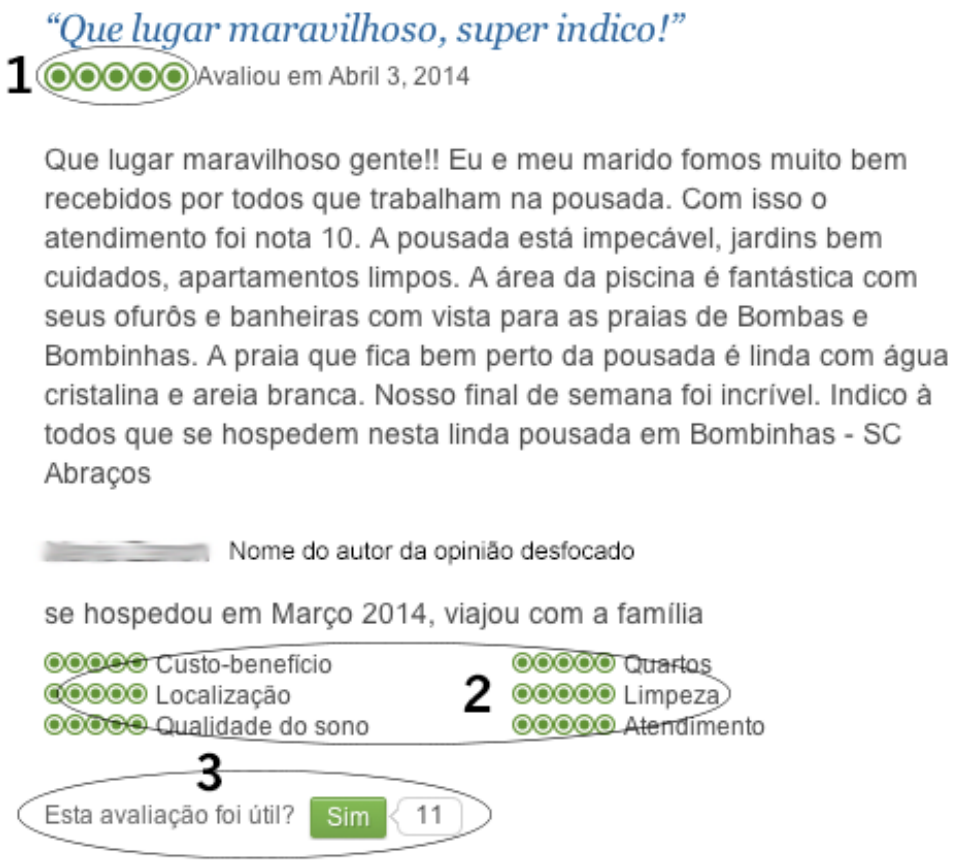

Figura 6: Descrição dos elementos de uma opinião

Fonte: TripAdvisor

sertação, uma versão da ferramenta Coh-Metrix-Port foi desenvolvida baseando-se na documentação disponível *online*[1], somente com as métricas básicas (conforme apresentado na tabela 5), pois o código original e funcional do Coh-Metrix-Port não estava disponível publicamente.

No entanto, As métricas básicas do Coh-Metrix-Port são semelhantes as utilizadas nos trabalhos de Cao et al. (2011) e Korfiatis et al. (2012). A partir dessa implementação simplificada do Coh-Metrix-Port, foi possível extrair as características textuais: índice de inteligibilidade (em uma escala de 0 a 100) e as contagens de sentenças, palavras, sílabas, média de palavras por sentenças, quantidade de verbos, substantivos, adjetivos, advérbios e pronomes.

[1]http://www.nilc.icmc.usp.br:3000/index/info

Tabela 5: Métricas do Coh-Metrix-Port utilizadas

Classe	Descrição
Índice de Inteligibilidade	Índice Flesch Reading Easy
Contagens Básicas	Número de sentenças Número de parágrafos Número de palavras por sentenças Número de sílabas por palavras Média de sílabas por palavras de conteúdo (substantivos, verbos, adjetivos e advérbios)
Frequências	Incidência de verbos Incidência de substantivos Incidência de adjetivos Incidência de advérbios Incidência de pronomes

Fonte: Autoria própria.

4.6 CARACTERÍSTICAS SEMÂNTICAS DAS OPINIÕES

As características semânticas resumem a ideia principal da opinião e para Cao et al. (2011) representam a principal influência no voto de utilidade. No entanto, a subjetividade das opiniões dificulta a avaliação exata do significado da opinião.

Dessa forma, Cao et al. (2011) utilizou a LSA para extrair as características semânticas de todas as opiniões utilizadas em seu trabalho e avaliá-las em uma abordagem mais geral, isto é, ao invés de investigar quais as palavras que possuem influência em cada opinião, o objetivo foi avaliar se o conjunto dessas características auxilia na percepção da utilidade das opiniões.

4.7 DEFINIÇÃO DOS MODELOS DE REGRESSÃO LOGÍSTICA ORDINAL

Para determinar como e quais características de uma opinião influenciam no número de votos de utilidade, foram criados cinco modelos *OLR* com combinações entre os três conjuntos de características básicas, textuais e semânticas seguindo o mesmo arranjo discutido no trabalho de Cao et al. (2011).

A pesquisa de Cao et al. (2011) utiliza arbitrariamente conjuntos de 50, 100, 150 e 200 colunas da matriz resultado da LSA a qual denominou de "Fatores SVD", pois afirma que seu estudo é uma prova de conceito da metodologia utilizada. Por esse motivo, apresenta somente o resultado de um experimento com 100 Fatores SVD, deixando em aberto a escolha de outra metodologia para seleção das variáveis.

No entanto, para repetir os mesmos tipos de avaliações dos modelos e para selecionar os Fatores SVD desta pesquisa, foi realizada uma seleção aleatória e sem repetição de 50, 100, 150 e 200 colunas da matriz LSA. Cada um dos conjuntos de colunas selecionados representa um cenário de experimentação. A mesma denominação de "Fatores SVD" será utilizada desse ponto em diante para referenciar as colunas da matriz LSA utilizadas neste trabalho.

A Tabela 6 apresenta a descrição dos modelos e a quantidade das variáveis independentes em cada um dos cenários avaliados. Observar que os Fatores SVD somente influenciam os modelos 3 e 5 que utilizam características semânticas. A *OLR* foi utilizada para investigar a relação entre as variáveis das três características da opinião e o número de votos de utilidade que uma opinião recebe.

A variável dependente deste estudo é um "ranking de utilidade" baseado no número de votos das opiniões (ver Tabela 3). Nesta pesquisa, o número de votos de utilidade das opiniões foram agrupados de forma que os valores de 0 à 6 representam estes valores neste intervalo, enquanto o valor 7 agrupa "7 ou mais votos" em uma mesma categoria.

Os três primeiros modelos (i.e. modelos 1, 2 e 3) serão utilizados para verificar a influência de cada característica nos votos de utilidade separadamente. O modelo 4 combina características básicas e textuais e o modelo 5 agrega os três tipos de características. A comparação dos modelos 4 e 5 permite verificar se o agrupamento de características diferentes representa alguma alteração na avaliação dos votos de utilidade das opiniões.

Tabela 6: Descrição dos modelos e o número de variáveis inicial

Modelo	**Conjuntos de características**	**Cenários**			
		50	100	150	200
1	Básicas	9	9	9	9
2	Textuais	10	10	10	10
3	Semânticas (Fatores SVD)	50	100	150	200
4	Básicas + Textuais	19	19	19	19
5	Básicas + Textuais + Semânticas	69	119	169	219

Fonte: Autoria própria.

A regressão OLR dos modelos 1, 2 e 4 foi realizada independentemente do processamento da LSA, pois as variáveis básicas e textuais são o resultado da etapa de pré-processamento (seção 5.2.1). A regressão OLR do modelo 3 foi realizada após a criação da matriz semântica (seção 5.2.3).

As variáveis utilizadas no modelo 3 foram separadas em conjuntos de 50, 100, 150

e 200 Fatores SVD, isto é, os 200 termos selecionados foram subdivididos em intervalos de 1 até 50, 1 até 100, 1 até 150 e finalmente, 1 até 200. Esta separação deu origem aos cenários de observação. O modelo 5 é o resultado do agrupamento das variáveis utilizadas originalmente nos modelos 3 e 4, inclusive seguindo a mesma distribuição de cenários descritos anteriormente.

No entanto, o modelo 5 obteve a maior quantidade de variáveis entre os modelos, o que pode aumentar o seu ajuste e diminuir sua capacidade de predição. Conforme descrito no trabalho de Cao et al. (2011), para selecionar as variáveis com melhor capacidade de predição, foi utilizado o método de seleção automático de variáveis descrito anteriormente na seção 2.5.

Após a redução dos termos, restaram no novo modelo 5 apenas 17, 39, 56 e 78 variáveis para serem analisadas respectivamente em seus cenários, conforme a Tabela 7.

Tabela 7: Novo arranjo de variáveis no modelo 5

Modelo	**Conjuntos de características**	**Cenários**			
		50	100	150	200
1	Básicas	9	9	9	9
2	Textuais	10	10	10	10
3	Semânticas (Fatores SVD)	50	100	150	200
4	Básicas + Textuais	19	19	19	19
5	Básicas + Textuais + Semânticas	17	39	56	78

Fonte: Autoria própria.

4.8 CONCLUSÃO

Este capítulo apresentou a descrição do método proposto para a realização da comparação dos modelos das características das opiniões a partir do trabalho de Cao et al. (2011) apresentando uma aplicação da metodologia descrita em seu artigo para verificar se as características semânticas apresentam a mesma influência em opiniões no domínio de serviços (hotelaria) e escritas em português.

Além disso, para comparar com a afirmação do artigo de Korfiatis et al. (2012) a métrica de inteligibilidade discutida anteriormente é utilizada para avaliar se são as características textuais que possuem essa influência.

O próximo capítulo apresenta o experimento criado nesta dissertação demonstrando alguns aspectos do processamento das opiniões e das métricas de comparação utilizadas.

5 EXPERIMENTO

5.1 INTRODUÇÃO

No capítulo anterior foi apresentado o método adotado para a realização dos experimentos através da descrição dos modelos que seriam avaliados em conformidade com a metodologia do trabalho de Cao et al. (2011).

Neste capítulo será apresentado o processamento das opiniões e as métricas utilizadas para a comparação dos modelos. Os dados para o experimento desta pesquisa foram coletados de opiniões reais dos serviços de hotéis do *site* de viagens TripAdvisor (utilizando sua versão em português), que indexa entre outras informações, as opiniões dos viajantes de diversas cidades em mais de 40 países em todo o mundo.

O capítulo está organizado da seguinte maneira: a seção 5.2 aborda o passo-a-passo do processamento das opiniões utilizadas e a seção 5.3 descreve as métricas de comparação e desempenho dos modelos.

5.2 PROCESSAMENTO DAS OPINIÕES

O conteúdo das opiniões exibidas na área pública do *site* foram capturados e organizados em arquivos no formato texto. Foram incluídas somente as opiniões no intervalo de 1/1/2014 a 1/8/2014 (incluindo os dois extremos), pois o objetivo foi abranger um período de tempo razoável antes e durante a Copa do Mundo de Futebol 2014 (realizada entre 12/6 e 13/7 daquele ano).

O processamento foi realizado utilizando a linguagem R (R Core Team, 2014) em um servidor Linux Debian 7, 64 bits com 120 GB de RAM, 2TB de disco e 24 núcleos (Intel Xeon E5-2420 1.90GHz). A ferramenta Apache OpenNLP (APACHE, 2011) em conjunto com algoritmos próprios foi utilizada para realizar o processamento do texto, classificação sintática e identificação de palavras e sentenças (*"part-of-speech"*).

O quadro abaixo apresenta as principais bibliotecas R utilizadas no processamento das opiniões:

Quadro 5.1: Principais bibliotecas

```
library(MASS) # Text mining
library(tm)   # Text mining
library(lsa)  # Latent Semantic Analysis
library(rminer)
library(ordinal)
library(caret)
```

Conforme ilustrado na Figura 7, o processamento das opiniões foi dividido em três etapas: (i) Pré-processamento do texto, (ii) Redução dos termos e (iii) Aplicação da LSA.

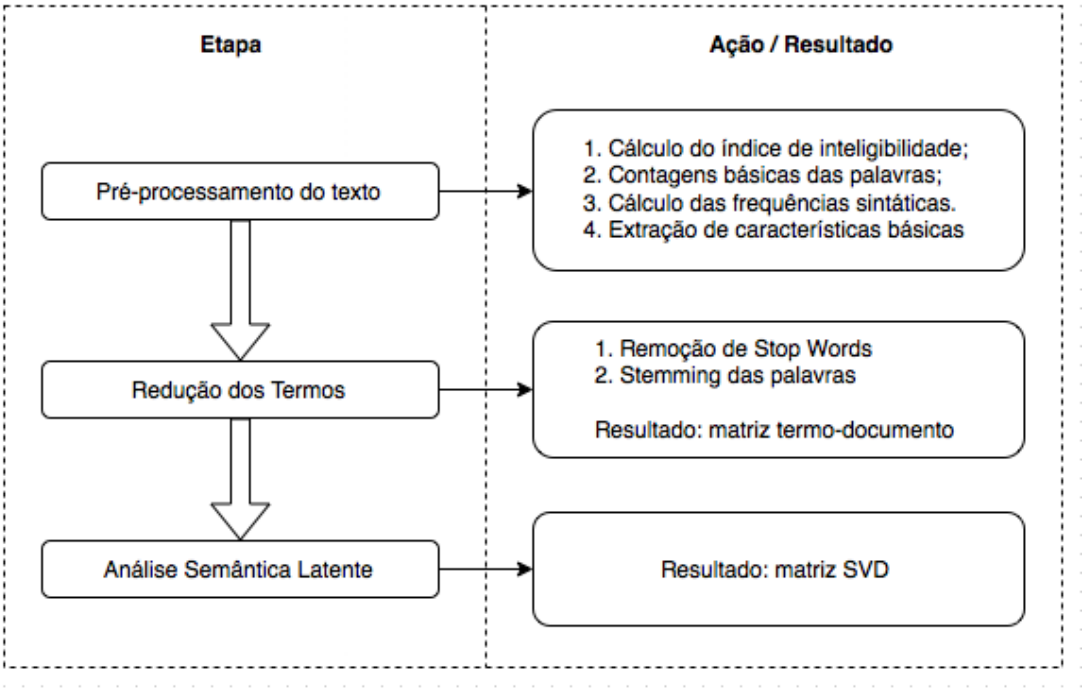

Figura 7: Etapas do processamento

Fonte: Autoria própria.

5.2.1 PRÉ-PROCESSAMENTO DO TEXTO

Na etapa de pré-processamento de cada opinião foi utilizada a versão simplificada do Coh-Metrix-Port (ver Tabela 5 acima) para calcular: índice de inteligibilidade, contagens básicas das palavras e as frequências sintáticas das palavras.

Também foi calculado o intervalo entre as datas da viagem do autor e da publicação da opinião, o nível de extremismo da opinião e por fim, as notas das características do hotel foram normalizadas atribuindo o valor 1 (um) para notas positivas e 0 (zero) para notas negativas (conforme a escala de valores discutida na seção 4.4).

O resultado desta etapa é uma matriz que possuí todas as variáveis necessárias para os modelos que utilizam as características básicas e textuais das opiniões.

5.2.2 REDUÇÃO DOS TERMOS

Na etapa seguinte, a redução dos termos, foram realizados a remoção de *stop words*[1] e o *stemming* das palavras, reduzindo a ambiguidade do texto para a próxima etapa. As opiniões filtradas são a entrada para a construção de uma matriz esparsa (conhecida com *Bag of Words* ou *BOW*) resultante do processamento da *TF-IDF*, onde cada coluna representa um documento e cada linha a frequência do termo.

O produto desta etapa é a matriz termo-documento. A Tabela 8 é um exemplo da matriz gerada.

Tabela 8: Exemplo da matriz termo-documento

Termo	**Opinião 1**	...	**Opinião**	...	**Opinião**	...	**Opinião**	...	**Opinião 35.037**
abacax	0	...	0	...	1	...	0	...	0
bonit	0	...	5	...	0	...	8	...	0
...	...	...	...	...	...	...	...	...	...
caf	0	...	5	...	0	...	8	...	0
charm	0	...	5	...	0	...	8	...	0
...	...	...	...	...	...	...	...	...	...
zweig	0	...	5	...	0	...	8	...	0

Fonte: Autoria própria.

5.2.3 APLICAÇÃO DA LSA

Na etapa da análise semântica latente, a matriz termo-documento foi utilizada como entrada para a LSA com o objetivo de reduzir a dimensionalidade da matriz e consequentemente extrair seu espaço semântico.

A matriz termos-documentos foi processada utilizando o função *lsa* (WILD, 2014) do R com o parâmetro de dimensões calculado pela função *dimcalc_share*, que retorna as n primeiras linhas da matriz LSA, cuja avaliação dos valores singulares das linhas é maior

[1]Exemplos: as, e, os, de, para, com, sem, foi, etc

ou igual ao índice de corte 0,5. O processamento da LSA é necessário para modelos 3 e 5, pois as variáveis básicas e textuais foram extraídas diretamente utilizando informações já disponíveis no conjunto das opiniões ou processadas pelo Coh-Metrix-Port.

Quadro 5.2: Processamento LSA

```
1 corpus <- Corpus(VectorSource(MATRIZ_DOCUMENTOS))
2 corpus <-tm_map(corpus, removeNumbers)
3 corpus <-tm_map(corpus, stemDocument, language = "portuguese")
4 corpus <-tm_map(corpus, stripWhitespace)
5
6 dtm <- TermDocumentMatrix(corpus, control =
7           list(weighting = function(x) weightTfIdf(x, normalize = FALSE)))
8 dtm_matrix <- as.matrix(dtm)
9 lsa_space <- lsa(dtm_matrix, dims=dimcalc_share())
```

5.3 MÉTRICAS DESEMPENHO E COMPARAÇÃO DOS MODELOS

Para manter a conformidade com o trabalho de Cao et al. (2011) na comparação dos resultados estatísticos entre os modelos, foram utilizadas as mesmas métricas de comparação de seu trabalho nos modelos do experimento desta pesquisa. Os códigos em linguagem R criados para realizar os cálculos das métricas serão apresentadas nas seções 5.3.1, 5.3.2 e 5.3.3.

5.3.1 TAXA DE CLASSIFICAÇÃO INCORRETA

A taxa de classificação incorreta é utilizada para calcular o desempenho de um classificador em função das classificações incorretas que faz. É calculado através da proporção de classificações incorretas sobre o total de classificações do modelo e quanto maior a taxa de classificação incorreta do modelo, pior é o seu desempenho. O critério utilizado no cálculo da taxa de classificação incorreta é:

$$MR = \frac{\#\text{ de casos incorretamente classificados}}{\text{número de casos}} \tag{10}$$

Para o cálculo da taxa de classificação incorreta dos cinco modelos OLR foi utilizado a função *mmetric* (CORTEZ, 2010) e a métrica "CE" (linha 2 do quadro 5.3). Essa função recebe como parâmetros de entrada a variável dependente do modelo e as probabilidades calculadas através da função *predict* (JR, 2014). A resposta do processamento da

função é o índice de classificação incorreto. Para a validação manual do índice, foi criada a matriz de confusão utilizando a mesma função *mmetric* e a métrica "CONF" (linha 1 do quadro 5.3).

Quadro 5.3: Taxa de Classificação incorreta

```
1 confusion_matrix <- mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("CONF"))
2 misclas_rating <- mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("CE"))
```

5.3.2 CRITÉRIO DE INFORMAÇÃO DE AKAIKE

O AIC é uma das métricas utilizadas na regressão logística ordinal seu cálculo faz parte do resultado da função *clm* da biblioteca *Ordinal* (CHRISTENSEN, 2014) do R. O primeiro parâmetro da *clm*, apresentada no Quadro 5.4, é uma fórmula composta pelos nomes das variáveis que a função deverá utilizar e que estão na matriz de dados do parâmetro "data". A fórmula de entrada utilizada nos modelos deste experimento é o "ranking da opinião" (dependente) e as variáveis independentes disponíveis em cada modelo em cada um dos cenários. O parâmetro "link" recebe a função de ligação que deverá ser utilizada na regressão. A opção "logit" representa a função do modelo de chances proporcional (seção 2.3.1)

Quadro 5.4: Regressão logistica ordinal

```
1 modelo <- clm(FORMULA_ENTRADA, data = DADOS, link=c("logit"))
2 AIC(modelo);
```

5.3.3 RAZÃO DE LIFT (*LIFT RATIO*)

A razão de lift é utilizada para medir o desempenho de predição de um modelo em um segmento aleatório da população. Compara a predição do modelo para uma amostra da população contra a predição do modelo com a população. O desempenho do modelo é melhor se apresentar um ganho de predição para a amostra em relação ao conjunto de dados que está sendo analisado.

Essa métrica foi calculada através da função *mmetric* (CORTEZ, 2010) utilizando a opção "LIFT" como parâmetro. O parâmetro "T" representa o ponto de corte para cada classe do ranking de votos. O resultado da Razão de *Lift* de cada classe é acumulado no final do processamento (linha 8 do Quadro 5.5) e representa a Razão de *Lift* geral do modelo avaliado.

Quadro 5.5: Razão de Lift

```
T1 = mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("LIFT"), T=1)
T2 = mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("LIFT"), T=2)
T3 = mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("LIFT"), T=3)
T4 = mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("LIFT"), T=4)
T5 = mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("LIFT"), T=5)
T6 = mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("LIFT"), T=6)
T7 = mmetric(VARIAVEL_DEPENDENTE,PROBAB,metric=c("LIFT"), T=7)
lift_ratio <- T1 + T2 + T3 + T4 + T5 + T6 + T7
```

5.4 CONCLUSÃO

Este capítulo apresentou uma visão geral sobre as principais funções R das bibliotecas utilizadas para o processamento dos textos e dos cálculos das métricas que foram extraídos dos modelos.

O próximo capítulo apresenta o resultado do experimento e seus quatro testes, além da descrição da comparação dos cinco modelos propostos.

6 ANÁLISE DOS RESULTADOS

6.1 INTRODUÇÃO

A análise dos resultados permite identificar o modelo que obteve o melhor desempenho estatístico e, portanto, quais variáveis influenciam no número de votos da opinião. Todos os passos descritos na metodologia do trabalho de Cao et al. (2011) foram aplicados nesta pesquisa e os valores encontrados foram comparados com o trabalho de referência.

Neste capítulo será apresentado o resultado da aplicação da metodologia e dos cálculos estatísticos utilizados para comparar os modelos, além da discussão do significado destes resultados.

O restante do capítulo está organizado da seguinte maneira: a seção 6.2 apresenta o resultado encontrado por Cao et al. (2011) para servir de base de comparação sobre o desempenho dos modelos e a seção 6.3 discute o resultado da comparação dos modelos de acordo com as métricas de comparação utilizadas.

6.2 VALORES DE REFERÊNCIA

A Tabela 9 é uma transcrição dos resultados da comparação dos cinco modelos OLR da pesquisa de Cao et al. (2011). O modelo destacado (negrito) combina as três características das opiniões e obteve o melhor resultado estatístico apresentando a menor taxa de classificação incorreta, o menor AIC e a maior Razão de *Lift*.

Segundo Cao et al. (2011), as características semânticas em conjunto com características básicas e textuais têm um importante papel na influência do número de votos de utilidade. Porém, sua conclusão deixa claro que uma das limitações de sua pesquisa foi analisar estas características semânticas de forma agrupada, ao invés de verificar especificamente quais palavras utilizadas nas opiniões influenciam no número de votos de utilidade.

Tabela 9: Resultado de Cao et al. (2011) na comparação dos modelos

Modelo	MR	AIC	Razão de *Lift*
Modelo 1	0.48301	9863.15	5.92
Modelo 2	0.48589	9920.73	4.32
Modelo 3	0.48560	9990.31	4.64
Modelo 4	0.48243	9792.87	5.76
Modelo 5	**0.47753**	**9599.26**	**7.36**

Fonte: (CAO et al., 2011)

A mesma avaliação das características semânticas de forma agrupada foi assumida para o experimento desta pesquisa. A Tabela 10 apresenta a variável e sua descrição para facilitar o entendimento da nomenclatura utilizada nas próximas seções.

Tabela 10: Descrição das variáveis

Variável	Descrição
interval_date	Intervalo publicação/viagem
level	Extremismo da avaliação
hotel_value	Custo-benefício
hotel_sleep_quality	Qualidade do sono
hotel_service	Atendimento
hotel_rooms	Quartos
hotel_cleanliness	Limpeza
hotel_location	Localização
fleschsc	Índice Flesch Reading Easy
numsente	Número de sentenças
nunwords	Número de palavras por sentenças
nunsylla	Número de sílabas por palavras
meanwsen	Média de sílabas por palavras de conteúdo
verb_cnt	Incidência de verbos
noun_cnt	Incidência de substantivos
adjc_cnt	Incidência de adjetivos
advb_cnt	Incidência de advérbios
pron_cnt	Incidência de pronomes
SVD_N	Termos da matriz resultante da LSA

Fonte: Autoria própria.

6.3 COMPARAÇÃO DOS CENÁRIOS

A pesquisa de Cao et al. (2011) foi utilizados como base de comparação desta pesquisa. Os resultados obtidos corroboram os resultados obtidos no trabalho de referência. Em cada cenário, o modelo 3 utiliza todo o seu conjunto de variáveis disponível, ou seja, no cenário 1 utiliza 50 Fatores SVD, no cenário 2 utiliza 100 Fatores SVD e assim por

diante. O modelo 5 é o único que utiliza as variáveis selecionadas pelo método de seleção automática passo a passo.

6.3.1 CENÁRIO 1

No cenário 1, a Tabela 11 destaca o resultado dos melhores valores das métricas de comparação. Apesar de ter o melhor desempenho no AIC e na Razão de *Lift*, o modelo 5 não apresentou um bom resultado na taxa de classificação incorreta.

O segundo melhor resultado para as métricas AIC e Razão de *Lift* deste cenário foi encontrado no modelo 4. A taxa de classificação incorreta foi semelhante nos modelos 1 e 4, sendo a menor observada no modelo 3.

No modelo 5 deste cenário foram utilizadas as variáveis básicas: "level", "interval_date", "hotel_service", "hotel_rooms", "hotel_cleanliness", "hotel_location"; as variáveis textuais: "nunsylla", "adjc_cnt"; e as variáveis semânticas: "SVD_1", "SVD_22", "SVD_24", "SVD_25", "SVD_29", "SVD_41", "SVD_43", "SVD_49", "SVD_50".

Embora o significado original não possa ser recuperado sem um processamento adicional, as variáveis semânticas acima poderiam ser respectivamente traduzidas como: "$$$$", "abast", "abat", "abdominal", "aberraco", "abord", "aborrec", "abracadinh" e "abram".

Tabela 11: Comparação dos modelos: Cenário 1

Modelo	**MR**	**AIC**	**Razão de *Lift***
Modelo 1	0.38105	77828.65	3.67
Modelo 2	0.38108	77692.44	3.68
Modelo 3	**0.38100**	77843.53	3.64
Modelo 4	0.38105	**77643.50**	**3.76**
Modelo 5	0.38111	**77582.25**	**3.77**

Fonte: Autoria própria.

6.3.2 CENÁRIO 2

No cenário 2, a Tabela 12 também destaca o melhor resultado do modelo 5 nas métricas AIC e Razão de *Lift*, porém, o mesmo não obteve um bom resultado na taxa de classificação incorreta. Similar ao cenário 1, o modelo 4 também obteve o segundo melhor resultado para a métrica AIC. No entanto, o modelo 3 apresentou o segundo melhor desempenho na taxa de classificação incorreta e na Razão de *Lift*.

O modelo 5 deste cenário utilizou as variáveis básicas: “level”, “interval_date”, “hotel_sleep_quality”, “hotel_service”, “hotel_cleanliness”, “hotel_location”; as variáveis textuais: “nunsylla”, “adjc_cnt”; e as variáveis semânticas: “SVD_3”, “SVD_5”, “SVD_8”, “SVD_11”, “SVD_12, “SVD_14”, “SVD_18”, “SVD_21”, “SVD_22”, “SVD_23”, “SVD_24, “SVD_25”, “SVD_26”, “SVD_30”, “SVD_34”, “SVD_44”, “SVD_47”, “SVD_53”, “SVD_61”, “SVD_62, “SVD_65, “SVD_66, “SVD_68”, “SVD_70”, “SVD_73”, “SVD_79”, “SVD_85”, “SVD_90”, “SVD_93”, “SVD_97, “SVD_99”.

Tabela 12: Comparação dos modelos: Cenário 2

Modelo	MR	AIC	Razão de *Lift*
Modelo 1	0.38105	77828.65	3.67
Modelo 2	0.38108	77692.44	3.68
Modelo 3	**0.38105**	77834.17	**3.78**
Modelo 4	0.38105	**77643.50**	3.76
Modelo 5	0.38108	**77530.67**	**3.84**

Fonte: Autoria própria.

6.3.3 CENÁRIO 3

No cenário 3, a Tabela 13 apresenta os melhores resultados nas métricas AIC e Razão de *Lift* para os modelos 4 e 5, porém, a melhor taxa de classificação incorreta foi obtida pelo modelo 3. Assim como no cenário 2, o modelo 3 também obteve um bom desempenho na Razão de *Lift*.

Tabela 13: Comparação dos modelos: Cenário 3

Modelo	MR	AIC	Razão de *Lift*
Modelo 1	0.38105	77828.65	3.67
Modelo 2	0.38108	77692.44	3.68
Modelo 3	**0.38077**	77789.43	**3.76**
Modelo 4	0.38105	**77643.50**	**3.76**
Modelo 5	0.38097	**77439.38**	**3.86**

Fonte: Autoria própria.

O modelo 5 deste cenário utilizou as variáveis básicas: “level”, “interval_date”, “hotel_service”, “hotel_rooms”, “hotel_cleanliness”, “hotel_location”; as variáveis textuais: “nunwords”, “adjc_cnt”, “advb_cnt”; e as variáveis semânticas: “SVD_2”, “SVD_6”, “SVD_10”, “SVD_11”, “SVD_18”, “SVD_19”, “SVD_20”, “SVD_22”, “SVD_25”, “SVD_26”, “SVD_30”, “SVD_34”, “SVD_36”, “SVD_40”, “SVD_41”, “SVD_44”, “SVD_45”, “SVD_47”, “SVD_54”, “SVD_57”, “SVD_58”, “SVD_62”, “SVD_67”, “SVD_68”, “SVD_73”, “SVD_77”,

"SVD_79", "SVD_83", "SVD_86", "SVD_87", "SVD_92", "SVD_104", "SVD_105", "SVD_110", "SVD_113", "SVD_114", "SVD_116", "SVD_120", "SVD_121", "SVD_122", "SVD_125", "SVD_131", "SVD_132", "SVD_135", "SVD_137", "SVD_149", "SVD_150".

6.3.4 CENÁRIO 4

Conforme apresentado na Tabela 14, o modelo 5 deste cenário alcançou o melhor resultados na taxa de classificação incorreta, no AIC e na Razão de *Lift*. O resultado obtido pelos modelos 3 e 4 acompanhou a tendência dos cenários anteriores. Porém neste cenário, o modelo 3 obteve o segundo melhor resultado na taxa de classificação incorreta e na Razão de *Lift*, enquanto o modelo 4 teve um bom desempenho somente no AIC.

O modelo 5 deste cenário utilizou as variáveis básicas: "level", "interval_date", "hotel_sleep_quality", "hotel_service", "hotel_cleanliness", "hotel_location"; as variáveis textuais: "nunsylla", "adjc_cnt", "advb_cnt"; e as variáveis semânticas: "SVD_1", "SVD_4", "SVD_6", "SVD_7", "SVD_10", "SVD_16", "SVD_18", "SVD_21", "SVD_24", "SVD_27", "SVD_30", "SVD_33", "SVD_37", "SVD_38", "SVD_39", "SVD_41", "SVD_46", "SVD_47", "SVD_49", "SVD_56", "SVD_62", "SVD_64", "SVD_70", "SVD_73", "SVD_74", "SVD_76", "SVD_77", "SVD_82", "SVD_88", "SVD_89", "SVD_95", "SVD_96", "SVD_100", "SVD_101", "SVD_102", "SVD_103", "SVD_105", "SVD_106", "SVD_107", "SVD_108", "SVD_115", "SVD_117", "SVD_121", "SVD_122", "SVD_125", "SVD_128", "SVD_129", "SVD_132", "SVD_136", "SVD_137", "SVD_138", "SVD_143", "SVD_144", "SVD_145", "SVD_152", "SVD_153", "SVD_159", "SVD_162", "SVD_168", "SVD_170", "SVD_172", "SVD_175", "SVD_178", "SVD_180", "SVD_188", "SVD_189", "SVD_191", "SVD_193", "SVD_194".

Tabela 14: Comparação dos modelos: Cenário 4

Modelo	MR	AIC	Razão de *Lift*
Modelo 1	0.38105	77828.65	3.67
Modelo 2	0.38108	77692.44	3.68
Modelo 3	**0.38057**	77721.34	**3.88**
Modelo 4	0.38105	**77643.50**	3.76
Modelo 5	**0.38048**	**77354.24**	**3.92**

Fonte: Autoria própria.

6.4 COMPARAÇÃO ENTRE OS MODELOS

Para facilitar a interpretação dos resultados esta seção apresenta os resultados das três métricas (taxa de classificação incorreta, AIC e Razão de *Lift*) e dos modelos

agrupados por cenários. Nos gráficos apresentados em seguida (Figuras 8, 9 e 10), a linha tracejada representa a comparação dos valores e a tendência do resultado.

6.4.1 CORRELAÇÃO ENTRE AS MÉTRICAS

Conforme apresentado na Tabela 15, a analise da correlação entre as métricas utilizando do índice de correlação de Pearson, é possível notar no Cenário 1 (seção 6.3.1) a forte influência do AIC sobre a taxa de classificação incorreta, onde a diminuição do AIC influenciou o aumento da taxa de erro. A Razão de *Lift* também é fortemente correlacionada com as outras duas métricas, onde o aumento do *Lift* diminui o AIC, porém, aumenta a taxa de classificação incorreta.

Tabela 15: Correlação entre as métricas: Cenário 1

Cenário	**MR**	**AIC**	**Razão de *Lift***
MR	1	-	-
AIC	-0,81	1	-
Razão de *Lift*	0,68	-0,92	1

Fonte: Autoria própria.

No entanto, no Cenário 2 (seção 6.3.2) a correlação entre variáveis é diferente do cenário anterior. A métrica AIC possui uma grande correlação com a taxa de classificação incorreta e com a Razão de *Lift*. Porém, esta última possui uma correlação fraca com a taxa de classificação incorreta. A correlação entre as métricas no cenário 2 é apresentado na Tabela 16.

Tabela 16: Correlação entre as métricas: Cenário 2

Cenário	**MR**	**AIC**	**Razão de *Lift***
MR	1	-	-
AIC	-0,67	1	-
Razão de *Lift*	0,18	-0,58	1

Fonte: Autoria própria.

A Tabela 17 apresenta a correlação no Cenário 3 (seção 6.3.3). A variável AIC possui uma correlação fraca em relação as outras variáveis e a Razão de *Lift* obteve o índice de correlação mais alto entre as métricas. A medida que a Razão de *Lift* aumenta, os valores de AIC diminuem. O mesmo resultado é possível notar na taxa de erro, porém, como a correlação não é muito forte, o Cenário 3 apresenta uma diminuição moderada para taxa de classificação incorreta a medida que os valores da Razão de *Lift* aumentam.

Tabela 17: Correlação entre as métricas: Cenário 3

Cenário	**MR**	**AIC**	**Razão de *Lift***
MR	1	-	-
AIC	-0,15	1	-
Razão de *Lift*	-0,39	-0,82	1

Fonte: Autoria própria.

O Cenário 4 (seção 6.3.4), conforme apresentado na Tabela 18, possui um índice de correlação forte em todas as variáveis. O AIC e a taxa de classificação incorreta possuem um crescimento aproximado, enquanto o aumento da Razão de *Lift* influência na diminuição de valores nas duas métricas.

Tabela 18: Correlação entre as métricas: Cenário 4

Cenário	**MR**	**AIC**	**Razão de *Lift***
MR	1	-	-
AIC	0,64	1	-
Razão de *Lift*	-0,96	-0,72	1

Fonte: Autoria própria.

6.4.2 TAXA DE CLASSIFICAÇÃO INCORRETA

A Figura 8 mostra que o modelo 3 apresenta a melhor taxa de classificação incorreta na maior parte dos cenários, somente perdendo para o modelo 5 no cenário 4. Como tendência, pode-se inferir que a utilização de todas as três características do texto da opinião melhoram a classificação dos modelos.

Obviamente, os modelos 1, 2 e 4 não sofrem variações nos diferentes cenários porque não são influenciados pelos Fatores SVD. Observar que o uso de características básicas e textuais isoladas ou em conjunto (modelos 1, 2 e 4) perdem para ambos modelos (3 e 5) que utilizam características semânticas nos cenários 3 e 4.

Por fim, mesmo apresentando melhor desempenho nos cenários 1, 2 e 3, o modelo 3 deve ser observado com ressalvas. Este resultado pode ter sido influenciado pela grande quantidade de variáveis deste modelo, o que pode ter causado um aumento no sobre ajuste do mesmo.

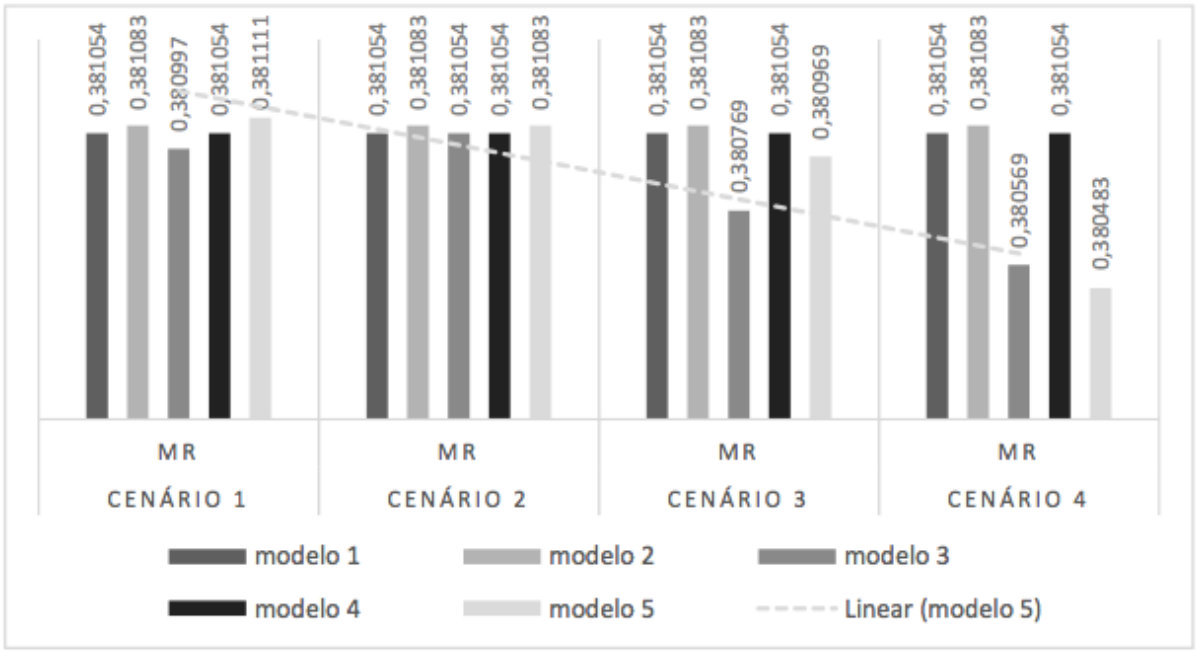

Figura 8: Comparação dos modelos: taxa de classificação incorreta
Fonte: Autoria própria.

6.4.3 AIC

A Figura 9 apresenta o resultado da comparação dos modelos nos 4 cenários utilizando a métrica AIC. Essa métrica foi extraída após o processamento da regressão OLR de cada modelo.

O modelo 5 foi o único pré-ajustado pela seleção automática de variáveis porque apresentava a maior quantidade de variáveis entre todos os modelos. O critério de inclusão de um previsor utilizado no método passo a passo também foi o valor AIC do modelo.

Neste caso, o modelo 5 apresentava previamente melhor AIC no momento do seu processamento pela regressão OLR. Por este motivo, é possível observar o melhor desempenho em todos os cenários (1 a 4).

Os modelos 1 e 3 apresentam os piores resultados em todos os cenários. Para o caso do modelo 3, pode-se inferir que a grande quantidade de variáveis está relacionado com o baixo desempenho. Esse resultado também confirma que as características textuais isoladas não produziram resultados significativos no ganho do modelo 1.

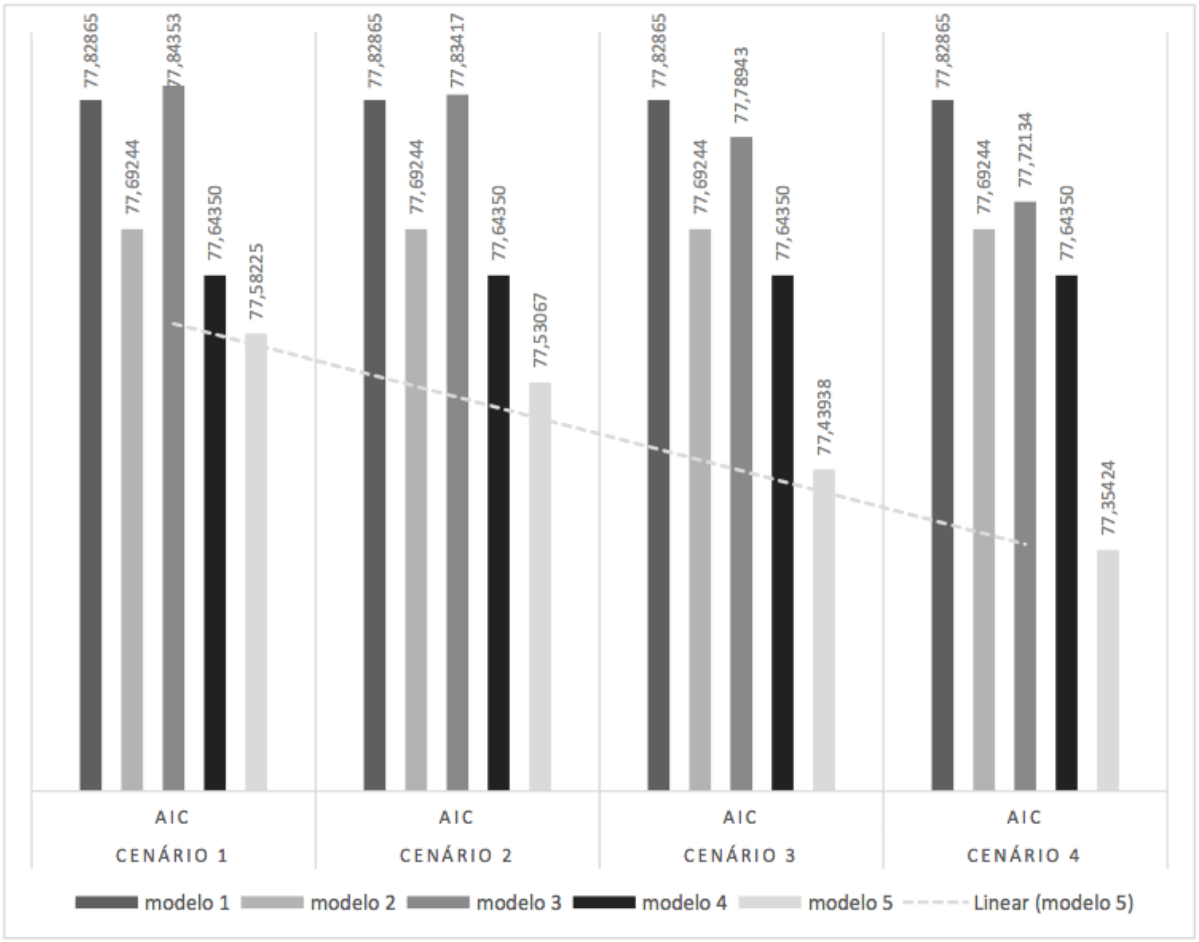

Figura 9: Comparação dos modelos: AIC

Fonte: Autoria própria.

6.4.4 RAZÃO DE *LIFT*

A figura 10 compara o resultado do ganho de predição alcançado pelos modelos nos 4 cenários avaliados e, em todos eles, o modelo 5 apresentou o melhor desempenho. Observar que a mesma influência da seleção automática das variáveis pode ser atribuído a esse resultado.

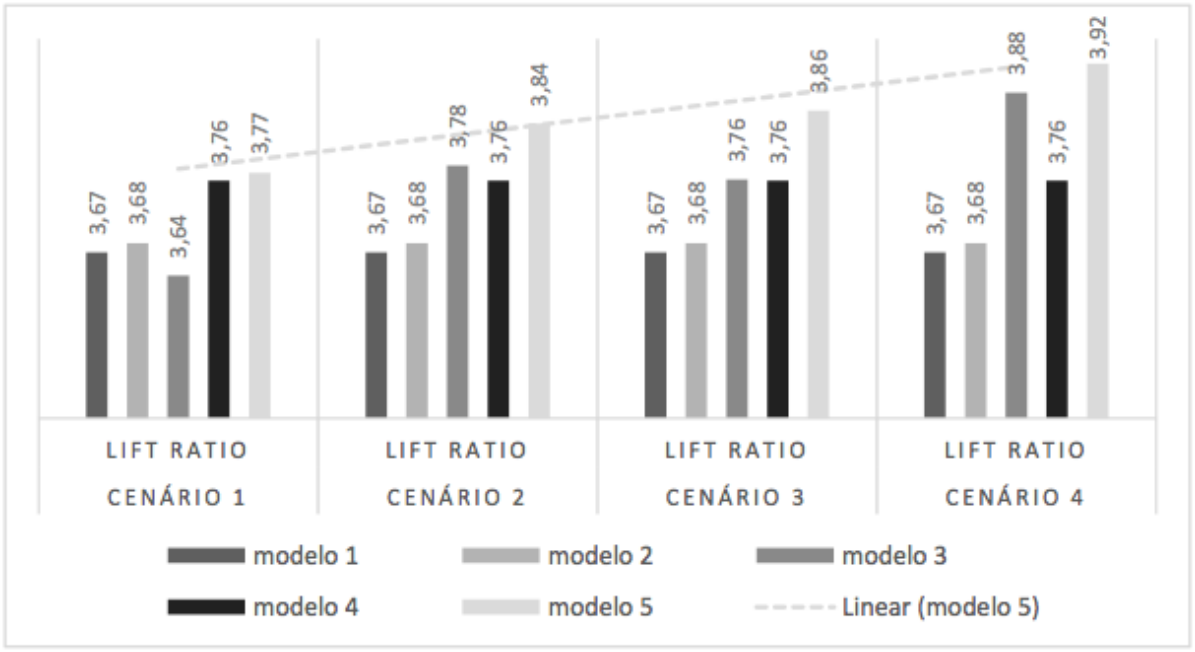

Figura 10: Comparação dos modelos: Razão de *Lift*
Fonte: Autoria própria.

Os modelos 1, 2 e 4 são os únicos que se mantêm estáveis em todos os cenários porque não possuem variação na quantidade de variáveis. A variação crescente da Razão de *Lift* do modelo 3, pode ser relacionada com a variação da quantidade de variáveis entre os cenários avaliados.

6.5 RESULTADO FINAL

A Tabela 19, descreve resumidamente o resultado final do modelo 5 (a tabela completa está anexada no apêndice A.3). Este modelo apresenta seis variáveis básicas (extremismo, intervalo de datas, qualidade do sono, atendimento, limpeza e localização), três variáveis textuais (número de sílabas, número de adjetivos e número de advérbios) e 69 variáveis semânticas.

Tabela 19: Resultado dos coeficientes do modelo

Variável	Coeficiente (desvio padrão)
level	0.041^{***} (0.012)
interval_date	-0.001^{***} (0.0003)
hotel_sleep_quality	-0.086^{**} (0.039)
hotel_service	-0.145^{***} (0.041)
hotel_cleanliness	-0.084^{**} (0.040)
hotel_location	0.152^{***} (0.041)
nunsylla	0.003^{***} (0.0003)
adjc_cnt	-0.018^{***} (0.004)
advb_cnt	-0.007 (0.005)
SVD_1	0.542^{*} (0.316)
SVD_4	-0.830^{**} (0.336)
SVD_6	-1.314^{**} (0.625)
...	...
SVD_176	0.044 (0.139)
SVD_177	0.043^{**} (0.021)
SVD_178	0.049^{***} (0.011)
SVD_179	0.199 (0.173)
SVD_180	0.393 (0.260)
SVD_181	-0.224 (0.362)
SVD_182	0.046^{**} (0.022)
SVD_183	0.011 (0.103)
SVD_184	-0.027 (0.049)
SVD_185	-0.093^{*} (0.053)
SVD_186	0.029 (0.029)
SVD_187	0.007 (0.017)
SVD_188	-0.584^{*} (0.336)
SVD_189	0.594 (0.494)
SVD_190	0.361 (0.479)
SVD_191	-0.008 (0.013)
SVD_192	-0.215 (0.645)
SVD_193	0.926 (0.677)
SVD_194	-0.036 (0.137)
SVD_195	-0.091 (0.072)
SVD_196	0.061 (0.058)
SVD_197	0.001 (0.025)
SVD_198	0.0003 (0.272)
SVD_199	0.254 (0.303)
SVD_200	-0.224 (0.243)
Log Likelihood	$-38{,}592.120$

Nota: *p<0.1; **p<0.05; ***p<0.01

Fonte: Autoria própria.

É importante notar que o P-valor de quatro variáveis básicas, duas variáveis textuais e 17 variáveis semânticas é menor que $0,01$, o que indica que são estatisticamente

significativas com 99% de nível de confiança. Outras duas variáveis básicas e 18 variáveis semânticas apresentaram o P-valor $0,05$ o que demonstra que são estatisticamente significativas com 95% de nível de confiança.

Ao contrário do trabalho de Cao et al. (2011), que chega a conclusão de que palavras de até quatro letras em inglês são mais simples de ler e podem significar mais votos para opinião, nesta pesquisa o resultado encontrado foi positivo para número de sílabas e negativo para as outras variáveis textuais, ou seja, quanto mais grupos de palavras curtas[1] e menos uso de adjetivos e advérbios, mais chances de uma opinião receber votos.

Além disso, diferente da conclusão de Korfiatis et al. (2012), o índice de inteligibilidade não foi incluído no modelo após a seleção automática de variáveis, o que significa que sua variação não representou melhora na capacidade de previsão do modelo em comparação com a utilização outras métricas textuais e características semânticas das opiniões.

O intervalo de publicação da opinião e a data efetiva da utilização do serviço é negativo, indicando que quanto menor esse intervalo, maior a chance da opinião receber mais votos, isto é, opiniões incluídas pouco tempo após a hospedagem no hotel podem ser consideradas mais úteis devido ao fato da experiência com o serviço ser recente. Uma opinião publicada muito tempo após a utilização do serviço pode não condizer com a realidade atual do serviço.

As opiniões "extremamente negativas" podem não exercer o mesmo efeito no domínio de serviços ou em outra cultura. Ao contrário do resultado encontrado em Cao et al. (2011), que observou que as pessoas tendem a votar nas opiniões mais negativas, a variável relacionada ao extremismo da opinião (*level*) do modelo 5, possui um valor positivo, ou seja, opiniões positivas sobre o serviço podem resultar em mais votos de utilidade.

Por possuir uma estimativa positiva, os usuários tendem a votar em opiniões com notas mais altas para "localização". Apresentando estimativas negativas, valores mais baixos para "qualidade do sono", "atendimento" e "limpeza" atraem mais votos, o que evidencia uma tendência de procura por serviços com conforto, higiene e bom atendimento. Estas quatro variáveis, podem indicar as características dos serviços de hotel que os usuários levam em consideração na escolha da empresa que irão contratar.

Finalmente, os coeficientes de alguns Fatores SVD do modelo 5 são positivos e

[1] A média de duas sílabas por palavras foi encontrada no conjunto total de opiniões analisadas

outros negativos. Isto indica que algumas palavras possuem influência positiva, atraindo mais votos e outras palavras possuem o efeito contrário, evitando que o usuário avalie a opinião.

A Figura 11 resume a comparação entre o resultado do modelo 5 e o resultado encontrado no trabalho de referência. Assim como no trabalho de Cao et al. (2011), o objetivo foi explorar se as características semânticas possuem influência no número de votos de utilidade e foi possível observar nos resultados empíricos da pesquisa que essas características efetivamente possuem impacto na avaliação de utilidade das opiniões também em língua portuguesa e no domínio de serviços de hotéis.

Grupos de variáveis	Variáveis	Modelo 5	(CAO et al., 2011)
Básicas	Extremismo da Opinião	Opiniões positivas sobre características atraem mais votos	Opinião negativa atrai mais votos
	Intervalo de publicação entre opinião e viagem	Pequenos intervalos influenciam positivamente	Pequenos intervalos influenciam positivamente
	Localização hotel	Muito importante	Não se aplica
	Limpeza	Importante	Não se aplica
	Qualidade do sono	Importante	Não se aplica
	Atendimento	Importante	Não se aplica
Textuais	Índice Flesch Reading Easy	Não teve influência	Não foi utilizada
	Sílabas por palavras	Palavras curtas	Até 4 letras
	Incidência de adjetivos	Poucos influenciam positivamente	Não foi utilizada
	Incidência de advérbios	Poucos influenciam positivamente	Não foi utilizada
Semânticas	Fatores SVD	Importante	Importante

Figura 11: Quadro comparativo dos resultados
Fonte: Autoria própria.

6.6 CONCLUSÃO

Este capítulo apresentou o resultado da aplicação da metodologia descritas na pesquisa de Cao et al. (2011) para encontrar quais características de um conjunto de opiniões possuí influência na percepção de utilidade de uma opinião e consequentemente na quantidade de votos de utilidade que uma opinião pode receber.

No próximo capítulo serão apresentadas as considerações finais, onde serão destacadas as contribuições da pesquisa e propostos trabalhos futuros.

7 CONCLUSÃO

Esta pesquisa teve o objetivo de adaptar e estender o trabalho de Cao et al. (2011) e assumiu algumas de suas limitações, como o número arbitrário de 200 Fatores SVD para verificar os passos necessários e a possibilidade da aplicação das mesmas técnicas em português. A seleção das métricas de comparação dos modelos e a metodologia de seleção de variáveis também é semelhante a utilizada no trabalho de referência.

Como uma das contribuições acadêmicas da presente dissertação, as informações das opiniões coletadas para o experimento da estão disponíveis publicamente[1], pois foram removidos quaisquer dados sensíveis e de identificação dos autores das opiniões.

As opiniões utilizadas são específicas para hotéis e possuem suas características próprias, como localização do hotel, detalhes referentes aos quartos, atendimento dos funcionários, entre outros. Por esse motivo, foi importante notar que notas mais altas para "localização" e valores mais baixos para "qualidade do sono", "atendimento" e "limpeza" atraem mais votos.

Curiosamente, a utilização do índice de inteligibilidade não trouxe nenhum ganho adicional no desempenho do modelo e nem mesmo foi incluída na seleção automática de variáveis do modelo. Porém, em um estudo preliminar utilizando uma segmentação diferente da mesma base de dados desta dissertação, o índice de inteligibilidade foi incluído entre as características textuais que influenciavam no voto de utilidade. Isto pode estar relacionado com a seleção automática das variáveis e do conjunto diferente de Fatores SVD utilizados.

Assim como Cao et al. (2011) percebeu que a quantidade de letras das palavras era importante, neste trabalho, o número de sílabas por palavras, a quantidade de adjetivos e advérbios das opiniões pode influenciar no número de votos. Entre as características básicas, opiniões positivas ou publicadas pouco tempo após a utilização do serviço são consideradas mais úteis do que opiniões antigas ou negativas.

[1] Ver endereço eletrônico: https://github.com/base16soft/mestrado

Foi percebida uma lacuna na comparação dos modelos e a necessidade de avaliar se o resultado poderia variar se a seleção automática das variáveis dos modelos fosse realizada em todos os modelos. Além disso o trabalho de Cao et al. (2011) avalia a variação do sinal obtido pela regressão OLR como um fator de influência da variável no texto da opinião, porém, não foram calculados quais as chances da alteração dos valores em uma variável aumentar ou diminuir o "ranking da opinião".

Foi observado que as características semânticas também são importantes nas opiniões no domínio de serviços, semelhante ao trabalho de Cao et al. (2011), que utilizou opiniões do domínio de *software* de computador, sendo este um domínio pouco explorado na literatura, visto que a maioria dos trabalhos encontrados, utilizam opiniões de produtos, livros, filmes e até mesmo hotéis.

Infelizmente, as características semânticas também representam uma dependência da base de dados utilizada no experimento, o que impediu a criação de um modelo genérico capaz de encontrar a melhor ordenação para novas opiniões. Essa dependência foi identificada, pois, para extrair as características semânticas é necessário aplicar a LSA e a seleção de variáveis. Somente após essas etapas é possível encontrar o conjunto de Fatores SVD que melhoram o desempenho do modelo.

Foi possível encontrar uma "diferença cultural", pois, segundo Cao et al. (2011), as opiniões negativas (em inglês) atraem a atenção dos leitores. No entanto, em português foi percebido o contrário, sendo avaliações positivas um dos critérios que aumentam a probabilidade de receber mais votos de utilidade. Com este trabalho foi possível chegar a resultados aproximados aos encontrados no trabalho de Cao et al. (2011), confirmando a sua conclusão sobre as características semânticas, estendendo sua pesquisa em outro domínio e linguagem.

No entanto, não foi determinado qual número de sílabas, qual o tempo mínimo de publicação e quais adjetivos obtiveram estes resultados, sendo encontrar o conjunto "recomendado" das palavras, uma sugestão de trabalho futuro. Além disso, quais são efetivamente as palavras que capturam a atenção dos leitores? Essa é uma pergunta pode levar a descoberta de padrões de textos que potencializam as chances de leitura e mais votos em uma determinada opinião.

Porém, exceto pela inclusão da etapa de cálculo da métrica de inteligibilidade utilizando uma versão da ferramenta Coh-Metrix-Port implementada pelo autor da dissertação, foram seguidas todas as etapas da metodologia de Cao et al. (2011), alterando o domínio das opiniões para o segmento de serviços de hotéis e a linguagem utilizando

opiniões escritas em português extraídas do site TripAdvisor.

Certamente estas são avaliações que não puderam ser respondidas nesta pesquisa de mestrado e servem de sugestão para trabalhos futuros, pois, entre as opiniões capturadas existem informações sobre Restaurantes e Pontos Turísticos das cidades escolhidas para a Copa do Mundo de 2014 sendo possível comparar se os resultados se mantêm para outros segmentos turísticos.

REFERÊNCIAS

AGRESTI, A. Inference for contingency tables. **Categorical Data Analysis, Second Edition**, Wiley Online Library, p. 70–114, 2002.

AGRESTI, A. **Analysis of Ordinal Categorical Data**. [S.l.]: Wiley, 2012. (Wiley Series in Probability and Statistics).

AKAIKE, H. A new look at the statistical model identification. **Automatic Control, IEEE Transactions on**, v. 19, n. 6, p. 716–723, Dec 1974.

ANDERSON, J. A. Regression and ordered categorical variables. **Journal of the Royal Statistical Society. Series B (Methodological)**, JSTOR, p. 1–30, 1984.

APACHE. **OpenNLP**. 2011.

BARBOZA, E. M. F.; NUNES, E. M. d. A. A inteligibilidade dos websites governamentais brasileiros e o acesso para usuários com baixo nível de escolaridade. **Inclusão Social**, v. 2, n. 2, p. 19–33, 2007.

CAO, Q.; DUAN, W.; GAN, Q. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. **Decision Support Systems**, Elsevier, v. 50, n. 2, p. 511–521, jan. 2011.

CHEN, C.-K.; HUGHES, J. J. Using ordinal regression model to analyze student satisfaction questionnaires. ir applications, volume 1, may 26, 2004. **Association for Institutional Research (NJ1)**, ERIC, 2004.

CHEN, H.; ZIMBRA, D. AI and opinion mining. **Intelligent Systems, IEEE**, v. 25, n. 3, p. 74–80, 2010.

CHEN, H.-N.; HUANG, C.-Y. An investigation into online reviewers' behavior. **European Journal of Marketing**, Emerald Group Publishing Limited, v. 47, n. 10, p. 1758–1773, 2013.

CHEN, P.-Y.; DHANASOBHON, S.; SMITH, M. D. **All Reviews are Not Created Equal: The Disaggregate Impact of Reviews and Reviewers at Amazon.com**. [S.l.]: SSRN eLibrary, 2008.

CHRISTENSEN, R. H. B. **ordinal—Regression Models for Ordinal Data**. 2014. R package version 2014.11-14 http://www.cran.r-project.org/package=ordinal/.

CORTEZ, P. Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool. In: PERNER, P. (Ed.). **Advances in Data Mining – Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining**. Berlin, Germany: LNAI 6171, Springer, 2010. p. 572–583.

CROSSLEY, S. A. et al. A linguistic analysis of simplified and authentic texts. **The Modern Language Journal**, Blackwell Publishing Inc, v. 91, n. 1, p. 15–30, 2007.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE**, v. 41, n. 6, p. 391–407, 1990.

DUBAY, W. H. The principles of readability. **Costa Mesa, CA: Impact Information**, 2004.

EKMAN, P. An argument for basic emotions. **Cognition & emotion**, Taylor & Francis, v. 6, n. 3-4, p. 169–200, 1992.

FIELD, A. **Descobrindo a estatistica usando o SPSS - 2.ed.:**. [S.l.]: Bookman, 2009.

GHOSE, A.; IPEIROTIS, P. G. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In: **Proceedings of the ninth international conference on Electronic commerce**. [S.l.]: ACM, 2007. p. 303–310.

GHOSE, A.; IPEIROTIS, P. G. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. **IEEE Transactions on Knowledge and Data Engineering**, v. 23, n. 10, p. 1498–1512, out. 2011.

GOLUB, G.; KAHAN, W. Calculating the singular values and pseudo-inverse of a matrix. **J. Soc. Indust. Appl. Math.: Ser. B, Numer. Anal.**, v. 2, p. 205–224, 1965.

GRAESSER, A. C.; MCNAMARA, D. S.; LOUWERSE, M. M. What do readers need to learn in order to process coherence relations in narrative and expository text? In: **Rethinking reading comprehension**. [S.l.]: Guilford, 2003. p. 82–98.

GRAESSER, A. C. et al. Coh-metrix: Analysis of text on cohesion and language. In: **M. Louwerse - Topics in Cognitive Science**. [S.l.: s.n.], 2004. p. 27.

HOCKING, R. R. A biometrics invited paper. the analysis and selection of variables in linear regression. **Biometrics**, JSTOR, p. 1–49, 1976.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, v. 28, n. 1, p. 11–21, 1972.

JONES, K. S. Idf term weighting and ir research lessons. **Journal of documentation**, Emerald Group Publishing Limited, v. 60, n. 5, p. 521–523, 2004.

JR, F. E. H. **rms: Regression Modeling Strategies**. [S.l.], 2014. R package version 4.2-1. Disponível em: <http://CRAN.R-project.org/package=rms>.

KIM, S.-M. et al. Automatically assessing review helpfulness. In: **Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing**. [S.l.]: Association for Computational Linguistics, 2006. p. 423–430.

KORFIATIS, N.; GARCIA-BARIOCANAL, E.; SANCHEZ-ALONSO, S. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. **Electronic Commerce Research and Applications**, v. 11, n. 3, p. 205–217, maio 2012.

LANDAUER, T. K.; DUTNAIS, S. T. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. **PSYCHOLOGICAL REVIEW**, v. 104, n. 2, p. 211–240, 1997.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. **Discourse processes**, Taylor & Francis, v. 25, n. 2-3, p. 259–284, 1998.

LEE, J. What makes people read an online review? the relative effects of posting time and helpfulness on review readership. **Cyberpsychology, Behavior, and Social Networking**, v. 16, n. 7, p. 529–535, jul. 2013.

LI, M. et al. Helpfulness of online product reviews as seen by consumers: Source and content features. **International Journal of Electronic Commerce**, Taylor & Francis, v. 17, n. 4, p. 101–136, 2013.

LIPSCHUTZ, S.; LIPSON, M. **Algebra Linear: Coleção Schaum**. [S.l.]: Bookman, 2009. (Coleção Schaum).

LIU, B. Sentiment analysis and subjectivity. **Handbook of natural language processing**, Chapman & Hall/CRC, v. 2, p. 627–666, 2010.

LIU, B. Sentiment analysis and opinion mining. **Synthesis Lectures on Human Language Technologies**, v. 5, n. 1, p. 1–167, 2012.

LIU, J. et al. Low-quality product review detection in opinion summarization. In: **EMNLP-CoNLL**. [S.l.: s.n.], 2007. p. 334–342.

LIU, Y. et al. Modeling and predicting the helpfulness of online reviews. In: . [S.l.]: IEEE, 2008. p. 443–452.

LU, Y. et al. Exploiting social context for review quality prediction. In: **Proceedings of the 19th international conference on World wide web**. [S.l.]: ACM, 2010. p. 691–700.

MARTINS, T. B. F. et al. Readability formulas applied to textbooks in brazilian portuguese. **ICMC Technical Report**, scielo, v. 28, p. 11, 1996.

MCCULLAGH, P. Regression models for ordinal data. **Journal of the royal statistical society. Series B (Methodological)**, JSTOR, p. 109–142, 1980.

MCNAMARA, D. S.; LOUWERSE, M. M.; GRAESSER, A. C. **Coh-Metrix: Automated cohesion and coherence scores to predict text readability and facilitate comprehension**. [S.l.], 2002.

O'Mahony, M. P.; SMYTH, B. Learning to recommend helpful hotel reviews. In: **Proceedings of the third ACM conference on Recommender systems**. [S.l.]: ACM, 2009. p. 305–308.

PANG, B.; LEE, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In: **Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics**. [S.l.]: Association for Computational Linguistics, 2005. p. 115–124.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. **Foundations and Trends® in Information Retrieval**, v. 2, n. 1-2, p. 1–135, jan. 2008.

R Core Team. **R: A Language and Environment for Statistical Computing**. [S.l.], 2014.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. **Information processing & management**, Elsevier, v. 24, n. 5, p. 513–523, 1988.

SCARTON, C.; GASPERIN, C.; ALUISIO, S. Revisiting the readability assessment of texts in portuguese. In: **Advances in Artificial Intelligence–IBERAMIA 2010**. [S.l.]: Springer, 2010. p. 306–315.

SCARTON, C. E.; ALUISIO, S. M. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. **Linguamática**, v. 2, n. 1, p. 45–61, 2010.

TALWAR, A.; JURCA, R.; FALTINGS, B. Understanding user behavior in online feedback reporting. In: **Proceedings of the 8th ACM conference on Electronic commerce**. [S.l.]: ACM, 2007. p. 134–142.

TANG, J. et al. Context-aware review helpfulness rating prediction. In: **Proceedings of the 7th ACM conference on Recommender systems**. [S.l.]: ACM, 2013. p. 1–8.

TSAPARAS, P.; NTOULAS, A.; TERZI, E. Selecting a comprehensive set of reviews. In: **Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining**. [S.l.]: ACM, 2011. p. 168–176.

TSUR, O.; RAPPOPORT, A. RevRank: a fully unsupervised algorithm for selecting the most helpful book reviews. In: . [S.l.: s.n.], 2009.

TSYTSARAU, M.; PALPANAS, T. Mining subjective data on the web. 2010.

WIEMER-HASTINGS, P.; WIEMER-HASTINGS, K.; GRAESSER, A. Latent semantic analysis. In: CITESEER. **Proceedings of the 16th international joint conference on Artificial intelligence**. [S.l.]: Morgan Kaufmann Publishers Inc., 2004. p. 1–14.

WILD, F. **lsa: Latent Semantic Analysis**. [S.l.], 2014. R package version 0.73. Disponível em: <http://CRAN.R-project.org/package=lsa>.

ZHANG, Z.; VARADARAJAN, B. Utility scoring of product reviews. In: **Proceedings of the 15th ACM international conference on Information and knowledge management**. [S.l.]: ACM, 2006. p. 51–57.

ANEXO A – RESULTADO DO EXPERIMENTO

A.1 COEFICIENTES DO MODELO 1, 2, 4

A Tabela 20, lista os coeficientes dos modelos 1, 2 e 4 gerados pela LSA. O valor respresentado entre parênteses apresenta o desvio padrão da variável. Pelo fato dos modelos possuírem variáveis diferentes em alguns casos, como nos modelos 1 e 2, as variáveis que não existem em um determinado modelo, foram representadas por um traço (-).

Tabela 20: Coeficientes dos modelos 1, 2 e 4

	Modelo 1	Modelo 2	Modelo 4
level	0.013 (0.012)	-	0.048^{***} (0.013)
interval_date	-0.001^{***} (0.0003)	-	-0.001^{***} (0.0003)
hotel_sleep_quality	-0.054 (0.041)	-	-0.054 (0.041)
hotel_service	-0.153^{***} (0.042)	-	-0.138^{***} (0.042)
hotel_value	-0.039 (0.040)	-	-0.022 (0.040)
hotel_rooms	-0.031 (0.041)	-	-0.052 (0.041)
hotel_cleanliness	-0.073^{*} (0.042)	-	-0.074^{*} (0.042)
hotel_location	0.188^{***} (0.042)	-	0.163^{***} (0.042)
fleschsc	-	0.001 (0.003)	0.001 (0.003)
numsente	-	-0.004 (0.010)	-0.008 (0.010)
nunwords	-	0.003 (0.005)	0.004 (0.005)
nunsylla	-	0.002 (0.002)	0.002 (0.002)
meanwsen	-	0.0003 (0.004)	-0.0001 (0.004)
meansypw	-	0.098 (0.237)	0.086 (0.237)
verb_cnt	-	0.005 (0.008)	0.006 (0.008)
noun_cnt	-	-0.008 (0.008)	-0.007 (0.008)
adjc_cnt	-	-0.016^{**} (0.008)	-0.014^{*} (0.008)

advb_cnt	-	−0.014 (0.009)	−0.013 (0.009)
pron_cnt	-	−0.010 (0.012)	−0.010 (0.012)
verb_inc	-	−0.0004 (0.0004)	−0.0004 (0.0004)
noun_inc	-	0.00004 (0.0004)	0.00002 (0.0004)
adjc_inc	-	−0.0003 (0.0003)	−0.0003 (0.0003)
advb_inc	-	0.0001 (0.0004)	0.0001 (0.0004)
pron_inc	-	0.001 (0.001)	0.001 (0.001)
Log Likelihood	−38,899.320	−38,823.220	−38,790.750

Nota: ± Significance: *p<0.1; Significance: **p<0.05; Significance: ***p<0.01

A.2 COEFICIENTES DO MODELO 3

A Tabela 21, lista os coeficientes do modelo 3 gerados pela LSA. O valor respresentado entre parênteses apresenta o desvio padrão da variável. Pelo fato do modelo possuír mais de uma representação, as variáveis que não existem em uma determinada variação, foram representadas por um traço (-).

Tabela 21: Coeficientes do modelo 3

	50 Fatores SVD	100 Fatores SVD	150 Fatores SVD	200 Fatores SVD
SVD_1	0.061*** (0.019)	0.076 (0.134)	0.210 (0.756)	0.299 (0.343)
SVD_2	−0.001 (0.018)	−0.125 (0.140)	−1.995* (1.090)	0.017 (0.323)
SVD_3	−0.081 (0.242)	0.563* (0.330)	0.002 (0.551)	0.073 (0.058)
SVD_4	0.046 (0.311)	0.740 (0.582)	−0.688 (1.142)	−0.764** (0.340)
SVD_5	−0.364 (0.491)	1.267*** (0.478)	0.263* (0.154)	−0.232 (0.317)
SVD_6	0.085 (0.792)	−0.338 (0.381)	0.381** (0.171)	−1.429** (0.647)
SVD_7	0.030 (0.414)	0.005 (0.018)	0.134 (0.130)	1.289*** (0.417)
SVD_8	−0.051 (0.075)	2.645* (1.430)	−0.021 (0.059)	−0.623 (0.567)
SVD_9	0.434 (0.285)	0.044 (0.074)	0.115 (0.268)	0.413 (0.347)
SVD_10	0.337* (0.204)	0.046 (0.223)	0.007 (0.044)	−1.307** (0.619)
SVD_11	−0.001 (0.025)	1.212 (0.995)	1.601*** (0.513)	0.100 (0.140)
SVD_12	0.050 (0.089)	−0.593*** (0.208)	0.037 (0.128)	0.032 (0.025)
SVD_13	0.021 (0.018)	0.006 (0.023)	0.057 (0.356)	−0.009 (0.162)
SVD_14	0.093 (1.959)	0.375** (0.181)	0.444*** (0.165)	0.374 (0.596)
SVD_15	0.332 (0.515)	0.319 (0.218)	0.027 (0.027)	0.041 (0.099)

SVD_16	0.033** (0.017)	−0.393 (0.572)	−0.088 (0.374)	−0.737*** (0.250)
SVD_17	−0.078 (0.075)	0.894** (0.446)	0.031 (0.300)	0.110* (0.065)
SVD_18	−0.508 (0.488)	−0.929 (0.675)	2.263*** (0.669)	0.354** (0.164)
SVD_19	0.026 (1.429)	0.053 (0.061)	−0.049 (0.036)	−0.024 (0.305)
SVD_20	−0.307 (0.432)	−0.111 (0.214)	−0.084** (0.038)	0.100 (0.071)
SVD_21	−0.372 (0.627)	−0.039* (0.023)	−0.214 (0.462)	0.220 (0.253)
SVD_22	1.142*** (0.406)	−1.011** (0.432)	0.086*** (0.027)	0.605 (0.715)
SVD_23	0.269 (0.224)	0.033** (0.013)	0.278 (0.416)	0.712 (0.547)
SVD_24	0.787* (0.477)	0.978** (0.387)	0.292 (0.310)	0.443** (0.194)
SVD_25	−1.518* (0.854)	0.871*** (0.283)	−0.485* (0.267)	0.330 (0.336)
SVD_26	−0.173 (0.221)	−1.186** (0.524)	−0.339** (0.154)	−0.007 (0.258)
SVD_27	−0.017 (0.348)	0.405 (0.657)	0.570 (0.355)	−0.672 (0.459)
SVD_28	−0.069 (0.110)	−0.133 (0.469)	0.275 (0.476)	3.564 (3.502)
SVD_29	−0.710 (0.638)	−0.453 (0.577)	−0.032 (0.099)	0.002 (0.016)
SVD_30	0.220 (0.345)	0.522* (0.299)	0.419* (0.215)	0.345* (0.198)
SVD_31	0.056*** (0.021)	0.142 (0.263)	−0.223 (0.328)	−0.047 (0.344)
SVD_32	−5.678 (6.796)	0.048 (0.132)	−0.185 (0.280)	0.034*** (0.013)
SVD_33	−0.087 (0.217)	0.014 (0.134)	−0.333 (0.692)	0.907** (0.359)
SVD_34	1.333 (1.032)	−0.356 (0.291)	0.427*** (0.144)	−0.390 (0.476)
SVD_35	0.112 (0.850)	−0.014 (0.026)	0.002 (0.024)	0.028 (0.083)
SVD_36	0.071 (0.264)	0.030 (0.447)	0.190** (0.076)	−0.002 (0.053)
SVD_37	−0.192 (0.341)	0.366 (0.470)	−0.179 (0.170)	0.183* (0.095)
SVD_38	0.210 (0.174)	−0.067 (0.449)	−0.256 (0.400)	0.151*** (0.050)
SVD_39	0.179 (0.142)	−0.299 (0.754)	0.027 (0.025)	−0.278 (0.191)
SVD_40	0.015 (0.107)	0.003 (0.028)	−1.419*** (0.509)	0.117 (0.126)
SVD_41	0.323*** (0.088)	−1.319 (1.028)	−0.102 (0.080)	−0.276 (0.320)
SVD_42	0.076 (0.058)	0.023** (0.011)	0.073 (0.115)	0.229 (0.670)
SVD_43	1.213* (0.665)	0.098 (0.154)	−0.0001 (0.043)	−0.183 (0.646)
SVD_44	0.044** (0.020)	0.566** (0.276)	2.107** (1.030)	−0.105 (0.164)
SVD_45	−0.386 (0.315)	0.020 (0.156)	−1.485* (0.870)	−0.134 (0.253)
SVD_46	0.197 (0.596)	0.416 (0.488)	−0.029 (0.055)	−0.046 (0.034)
SVD_47	0.200 (0.241)	1.796*** (0.498)	0.068*** (0.015)	−1.444*** (0.424)
SVD_48	0.178 (0.192)	−0.071 (0.261)	−0.873 (1.825)	0.129 (0.401)
SVD_49	0.316*** (0.074)	−0.823 (0.744)	0.660 (0.409)	2.169*** (0.731)
SVD_50	0.057*** (0.009)	−0.034 (0.140)	0.076 (0.050)	−0.102 (0.138)

SVD_51	-	0.088 (0.143)	−0.009 (0.018)	0.003 (0.061)
SVD_52	-	0.259 (0.548)	−0.782 (0.935)	0.760* (0.451)
SVD_53	-	−0.544 (0.449)	0.037 (0.024)	−0.200 (0.312)
SVD_54	-	0.088* (0.050)	−0.031 (0.026)	−0.010 (0.021)
SVD_55	-	0.131 (0.111)	−0.237 (0.226)	−0.066 (0.134)
SVD_56	-	0.400 (0.673)	0.093 (0.080)	−0.022 (0.015)
SVD_57	-	0.083 (0.096)	−0.244 (0.161)	0.112 (0.298)
SVD_58	-	0.290 (0.393)	0.077** (0.034)	0.041 (0.187)
SVD_59	-	−0.511 (0.954)	0.092 (0.203)	0.095 (0.225)
SVD_60	-	0.111 (0.130)	0.094 (0.135)	−0.057 (0.244)
SVD_61	-	0.071** (0.028)	0.894 (0.903)	−0.748 (0.671)
SVD_62	-	−2.179** (0.856)	−6.697*** (1.904)	0.566*** (0.159)
SVD_63	-	−0.095 (0.215)	0.033 (0.586)	0.017 (0.025)
SVD_64	-	0.015 (0.497)	0.387 (0.443)	0.103*** (0.036)
SVD_65	-	0.753** (0.308)	0.197 (0.194)	0.214 (0.463)
SVD_66	-	−0.290 (0.243)	−0.240 (0.473)	−0.078 (0.438)
SVD_67	-	0.018 (0.027)	−2.050*** (0.505)	0.026 (0.057)
SVD_68	-	−0.549 (0.384)	0.623** (0.257)	0.001 (0.159)
SVD_69	-	−0.205 (0.165)	0.699 (0.751)	0.360 (0.468)
SVD_70	-	0.056*** (0.009)	0.042 (0.116)	0.428 (0.453)
SVD_71	-	0.307 (0.249)	0.010 (0.019)	0.084* (0.047)
SVD_72	-	0.114 (0.110)	−0.107 (0.177)	0.126 (0.295)
SVD_73	-	−0.372* (0.221)	−0.274*** (0.103)	−0.045 (0.046)
SVD_74	-	−0.482 (0.637)	0.012 (0.027)	−2.034*** (0.729)
SVD_75	-	0.134 (0.139)	0.087 (0.260)	0.203 (0.502)
SVD_76	-	-	0.020 (0.360)	2.031*** (0.646)
SVD_77	-	−0.216 (0.390)	0.050*** (0.013)	−0.541** (0.259)
SVD_78	-	−0.001 (0.031)	0.025 (0.019)	0.020 (0.012)
SVD_79	-	0.281*** (0.080)	0.180*** (0.042)	−0.135 (0.346)
SVD_80	-	−0.157 (0.186)	0.043 (0.029)	0.035 (0.474)
SVD_81	-	−0.524 (0.530)	0.125 (0.404)	−0.510 (0.357)
SVD_82	-	0.082 (0.422)	0.174 (0.152)	−1.378** (0.593)
SVD_83	-	−0.226 (0.243)	−0.699 (0.453)	−0.638 (0.661)
SVD_84	-	0.111 (0.307)	−0.159 (0.193)	−0.082 (0.230)
SVD_85	-	−1.402** (0.657)	0.188 (1.013)	0.025 (0.332)

SVD_86	-	0.094 (0.067)	−0.033*** (0.010)	0.493 (0.406)
SVD_87	-	0.022 (0.035)	0.269 (0.198)	0.008 (0.100)
SVD_88	-	0.040 (0.031)	1.412 (2.054)	0.182*** (0.060)
SVD_89	-	−0.107 (0.131)	−0.189 (0.344)	0.417** (0.207)
SVD_90	-	−1.232** (0.615)	−0.788 (0.556)	0.093 (0.098)
SVD_91	-	0.154 (0.136)	0.072 (0.046)	0.057 (0.210)
SVD_92	-	−0.052 (0.199)	−0.394** (0.165)	−0.209 (0.578)
SVD_93	-	0.288* (0.173)	0.435 (0.293)	0.206 (0.364)
SVD_94	-	−0.098 (0.189)	−0.098 (0.516)	0.265 (0.511)
SVD_95	-	0.132 (0.150)	0.035 (0.388)	−0.554** (0.275)
SVD_96	-	−0.004 (0.053)	0.040 (0.043)	−0.383* (0.231)
SVD_97	-	−0.937* (0.546)	−0.123 (0.267)	0.042 (0.426)
SVD_98	-	0.025** (0.011)	0.055 (0.412)	0.192 (0.467)
SVD_99	-	0.377*** (0.143)	−0.314 (0.380)	−0.103 (0.471)
SVD_100	-	−0.018 (0.030)	−0.017 (0.037)	−6.473*** (2.112)
SVD_101	-	-	0.024 (0.417)	0.199*** (0.069)
SVD_102	-	-	0.153 (0.158)	0.081*** (0.029)
SVD_103	-	-	1.038 (0.637)	−0.010 (0.042)
SVD_104	-	-	0.646*** (0.161)	0.726 (0.639)
SVD_105	-	-	0.047 (0.031)	1.887*** (0.437)
SVD_106	-	-	0.027 (0.037)	−0.048 (0.053)
SVD_107	-	-	1.032 (0.908)	0.227** (0.105)
SVD_108	-	-	0.102 (0.075)	0.071** (0.028)
SVD_109	-	-	0.144 (0.172)	0.154 (0.382)
SVD_110	-	-	0.899*** (0.340)	0.242* (0.130)
SVD_111	-	-	0.154 (0.669)	−0.043 (0.443)
SVD_112	-	-	−0.009 (0.053)	0.047 (0.182)
SVD_113	-	-	−0.527* (0.277)	−0.303 (0.434)
SVD_114	-	-	−9.307*** (2.256)	0.328 (0.340)
SVD_115	-	-	0.182 (0.299)	−1.257** (0.519)
SVD_116	-	-	−0.116 (0.135)	−0.061 (0.243)
SVD_117	-	-	0.009 (0.032)	0.382* (0.199)
SVD_118	-	-	0.099* (0.060)	0.049** (0.020)
SVD_119	-	-	0.380 (0.320)	0.021 (0.062)
SVD_120	-	-	−0.293 (0.184)	−0.367 (0.306)

SVD_121	-	-	-0.457^{*} (0.249)	-0.088^{***} (0.012)
SVD_122	-	-	1.061^{***} (0.312)	-0.015 (0.011)
SVD_123	-	-	-1.239 (1.412)	0.127 (0.104)
SVD_124	-	-	-0.216 (0.872)	-0.047 (0.263)
SVD_125	-	-	-0.051^{**} (0.023)	-0.731^{*} (0.443)
SVD_126	-	-	0.034 (0.026)	0.213 (0.306)
SVD_127	-	-	-0.178 (0.813)	0.179 (0.197)
SVD_128	-	-	0.107 (0.077)	1.150^{***} (0.336)
SVD_129	-	-	0.106 (0.066)	0.649^{***} (0.228)
SVD_130	-	-	0.073 (0.123)	0.218 (0.256)
SVD_131	-	-	-0.151 (0.289)	0.001 (0.024)
SVD_132	-	-	-0.108^{*} (0.056)	-0.819^{*} (0.449)
SVD_133	-	-	0.081 (0.058)	0.043^{***} (0.012)
SVD_134	-	-	-0.798 (0.530)	0.281 (0.291)
SVD_135	-	-	-0.717 (0.531)	-0.261 (0.652)
SVD_136	-	-	-0.030 (0.035)	1.529^{**} (0.747)
SVD_137	-	-	0.272^{**} (0.112)	-0.821^{**} (0.377)
SVD_138	-	-	-0.454 (0.578)	0.408^{**} (0.189)
SVD_139	-	-	-0.076 (0.721)	0.399 (0.409)
SVD_140	-	-	-0.076 (0.096)	-0.022 (0.060)
SVD_141	-	-	0.023 (0.038)	0.221^{*} (0.133)
SVD_142	-	-	-0.018 (0.035)	0.163 (1.247)
SVD_143	-	-	-0.037 (0.537)	-0.516 (0.456)
SVD_144	-	-	-0.033 (0.229)	0.980^{*} (0.506)
SVD_145	-	-	-0.124 (0.347)	0.725 (0.582)
SVD_146	-	-	0.133 (0.234)	0.441 (0.453)
SVD_147	-	-	-0.122 (0.236)	-0.320 (0.471)
SVD_148	-	-	0.470 (0.288)	0.045 (0.292)
SVD_149	-	-	-1.181^{**} (0.496)	-0.249 (0.394)
SVD_150	-	-	0.778^{*} (0.399)	0.311 (0.350)
SVD_151	-	-	-	0.725 (0.877)
SVD_152	-	-	-	0.148^{**} (0.061)
SVD_153	-	-	-	0.774^{*} (0.419)
SVD_154	-	-	-	0.645 (0.619)
SVD_155	-	-	-	0.093 (0.356)

SVD_156	-	-	-	-
SVD_157	-	-	-	0.023 (0.016)
SVD_158	-	-	-	0.147 (0.199)
SVD_159	-	-	-	-0.760^{*} (0.398)
SVD_160	-	-	-	0.0002 (0.211)
SVD_161	-	-	-	0.024 (0.161)
SVD_162	-	-	-	0.728^{**} (0.300)
SVD_163	-	-	-	-0.934 (0.624)
SVD_164	-	-	-	0.409 (0.337)
SVD_165	-	-	-	0.215 (0.648)
SVD_166	-	-	-	0.828 (0.665)
SVD_167	-	-	-	-1.848^{*} (1.011)
SVD_168	-	-	-	1.985^{***} (0.622)
SVD_169	-	-	-	-0.108 (0.151)
SVD_170	-	-	-	-0.089 (0.061)
SVD_171	-	-	-	0.002 (0.034)
SVD_172	-	-	-	-0.092^{**} (0.041)
SVD_173	-	-	-	-0.260 (0.366)
SVD_174	-	-	-	-0.001 (0.100)
SVD_175	-	-	-	-1.195 (0.788)
SVD_176	-	-	-	0.044 (0.139)
SVD_177	-	-	-	0.043^{**} (0.021)
SVD_178	-	-	-	0.049^{***} (0.011)
SVD_179	-	-	-	0.199 (0.173)
SVD_180	-	-	-	0.393 (0.260)
SVD_181	-	-	-	-0.224 (0.362)
SVD_182	-	-	-	0.046^{**} (0.022)
SVD_183	-	-	-	0.011 (0.103)
SVD_184	-	-	-	-0.027 (0.049)
SVD_185	-	-	-	-0.093^{*} (0.053)
SVD_186	-	-	-	0.029 (0.029)
SVD_187	-	-	-	0.007 (0.017)
SVD_188	-	-	-	-0.584^{*} (0.336)
SVD_189	-	-	-	0.594 (0.494)
SVD_190	-	-	-	0.361 (0.479)

SVD_191	-	-	-	−0.008 (0.013)
SVD_192	-	-	-	−0.215 (0.645)
SVD_193	-	-	-	0.926 (0.677)
SVD_194	-	-	-	−0.036 (0.137)
SVD_195	-	-	-	−0.091 (0.072)
SVD_196	-	-	-	0.061 (0.058)
SVD_197	-	-	-	0.001 (0.025)
SVD_198	-	-	-	0.0003 (0.272)
SVD_199	-	-	-	0.254 (0.303)
SVD_200	-	-	-	−0.224 (0.243)
Log Likelihood	−38,864.770	−38,811.080	−38,737.720	−38,654.670

Nota: ± Significance: *p<0.1; Significance: **p<0.05; Significance: ***p<0.01

A.3 COEFICIENTES DO MODELO 5

A Tabela 22, lista os coeficientes do modelo 5 gerados pela LSA após o processamento passo a passo para reduzir a quantidade de variáveis necessárias no modelo.

O valor respresentado entre parênteses apresenta o desvio padrão da variável. Pelo fato do modelo possuír mais de uma representação, as variáveis que não existem em uma determinada variação, foram representadas por um traço (-).

Tabela 22: Coeficientes do modelo 5

	17 Fatores SVD	39 Fatores SVD	56 Fatores SVD	78 Fatores SVD
level	0.039*** (0.012)	0.034*** (0.012)	0.050*** (0.012)	0.041*** (0.012)
interval_date	−0.001*** (0.0003)	−0.001*** (0.0003)	−0.001*** (0.0003)	−0.001*** (0.0003)
hotel_sleep_quality	-	−0.077** (0.039)	-	−0.086** (0.039)
hotel_service	−0.150*** (0.041)	−0.147*** (0.041)	−0.150*** (0.041)	−0.145*** (0.041)
hotel_rooms	−0.074* (0.038)	-	−0.072* (0.039)	-
hotel_cleanliness	−0.096** (0.040)	−0.094** (0.040)	−0.091** (0.040)	−0.084** (0.040)
hotel_location	0.150*** (0.040)	0.152*** (0.041)	0.147*** (0.040)	0.152*** (0.041)
nunsylla	0.002*** (0.0002)	0.002*** (0.0002)	-	0.003*** (0.0003)
nunwords	-	-	0.005*** (0.001)	-
adjc_cnt	−0.018*** (0.004)	−0.018*** (0.004)	−0.017*** (0.004)	−0.018*** (0.004)

advb_cnt	-	-	−0.009* (0.005)	−0.007 (0.005)
SVD_1	0.042** (0.019)	-	-	0.542* (0.316)
SVD_3	-	0.585* (0.319)	-	-
SVD_5	-	1.214*** (0.468)	-	-
SVD_8	-	2.725* (1.410)	-	-
SVD_4	-	-	-	−0.830** (0.336)
SVD_2	-	-	−2.240** (1.051)	-
SVD_6	-	-	0.294* (0.165)	−1.314** (0.625)
SVD_7	-	-	-	1.306*** (0.396)
SVD_10	-	-	−0.059 (0.044)	−1.398** (0.606)
SVD_11	-	1.413 (0.980)	1.493*** (0.497)	-
SVD_12	-	−0.647*** (0.200)	-	-
SVD_14	-	0.393** (0.176)	-	-
SVD_16	-	-	-	−0.771*** (0.243)
SVD_18	-	−1.076 (0.664)	2.302*** (0.659)	0.303* (0.159)
SVD_21	-	−0.042* (0.023)	-	0.335 (0.226)
SVD_19	-	-	−0.050 (0.036)	-
SVD_20	-	-	−0.085** (0.035)	-
SVD_22	1.178*** (0.386)	−0.973** (0.420)	0.062** (0.026)	-
SVD_23	-	0.029** (0.013)	-	-
SVD_24	0.937** (0.461)	1.000*** (0.384)	-	0.350* (0.190)
SVD_25	−1.385* (0.840)	0.763*** (0.275)	−0.391 (0.258)	-
SVD_29	−0.865 (0.612)	-	-	-
SVD_27	-	-	-	−0.728* (0.441)
SVD_41	0.307*** (0.087)	-	−0.138* (0.078)	−0.432 (0.307)
SVD_43	1.164* (0.662)	-	-	
SVD_46	-	-	-	−0.055 (0.034)
SVD_49	0.288*** (0.073)	-	-	2.182*** (0.684)
SVD_50	0.029*** (0.010)	-	-	-
SVD_26	-	−0.964* (0.513)	−0.473*** (0.151)	-
SVD_56	-	-	-	−0.030** (0.015)
SVD_30	-	0.448 (0.294)	0.339 (0.210)	0.282 (0.195)
SVD_34	-	−0.414 (0.285)	0.347** (0.141)	-
SVD_36	-	-	0.149** (0.074)	-
SVD_40	-	-	−1.225** (0.497)	-

SVD_44	-	0.542** (0.259)	1.624 (0.993)	-
SVD_45	-	-	−1.522* (0.844)	-
SVD_33	-	-	-	0.888** (0.352)
SVD_37	-	-	-	0.143 (0.092)
SVD_38	-	-	-	0.137*** (0.048)
SVD_39	-	-	-	−0.331* (0.182)
SVD_47	-	1.632*** (0.478)	0.048*** (0.015)	−1.340*** (0.412)
SVD_53	-	−0.716 (0.440)	-	-
SVD_61	-	0.067** (0.028)	-	-
SVD_54	-	-	−0.038 (0.026)	-
SVD_57	-	-	−0.287* (0.154)	-
SVD_58	-	-	0.075** (0.029)	-
SVD_62	-	−2.227*** (0.828)	−5.817*** (1.811)	0.441*** (0.159)
SVD_65	-	0.713** (0.303)	-	-
SVD_66	-	−0.328 (0.240)	-	-
SVD_67	-	-	−2.023*** (0.492)	-
SVD_68	-	−0.605 (0.371)	0.699*** (0.249)	-
SVD_64	-	-	-	0.056 (0.036)
SVD_70	-	0.028*** (0.010)	-	0.835** (0.419)
SVD_73	-	−0.404* (0.211)	−0.296*** (0.098)	−0.089** (0.045)
SVD_74	-	-	-	−2.055*** (0.712)
SVD_76	-	-	-	1.889*** (0.632)
SVD_77	-	-	0.040*** (0.012)	−0.560** (0.248)
SVD_79	-	0.251*** (0.077)	0.142*** (0.042)	-
SVD_85	-	−1.358** (0.649)	-	-
SVD_90	-	−1.045* (0.599)	-	-
SVD_93	-	0.277* (0.164)	-	-
SVD_97	-	−1.016* (0.534)	-	-
SVD_99	-	0.257** (0.120)	-	-
SVD_83	-	-	−0.619 (0.436)	-
SVD_86	-	-	−0.030*** (0.009)	-
SVD_87	-	-	0.274 (0.193)	-
SVD_92	-	-	−0.372** (0.155)	-
SVD_104	-	-	0.602*** (0.157)	-
SVD_82	-	-	-	−1.580*** (0.579)

SVD_88	-	-	-	0.161^{***} (0.059)
SVD_89	-	-	-	0.356^{*} (0.199)
SVD_95	-	-	-	-0.649^{**} (0.268)
SVD_96	-	-	-	-0.404^{*} (0.224)
SVD_100	-	-	-	-4.849^{**} (2.038)
SVD_101	-	-	-	0.125^{*} (0.066)
SVD_102	-	-	-	0.056^{*} (0.029)
SVD_103	-	-	-	-0.073^{*} (0.042)
SVD_105	-	-	0.049 (0.031)	1.793^{***} (0.430)
SVD_110	-	-	0.939^{***} (0.331)	-
SVD_113	-	-	-0.593^{**} (0.267)	-
SVD_114	-	-	-8.714^{***} (2.203)	-
SVD_116	-	-	-0.253^{*} (0.129)	-
SVD_120	-	-	-0.304^{*} (0.177)	-
SVD_106	-	-	-	-0.093^{*} (0.051)
SVD_107	-	-	-	0.153 (0.102)
SVD_108	-	-	-	0.057^{**} (0.028)
SVD_115	-	-	-	-1.004^{**} (0.483)
SVD_117	-	-	-	0.470^{**} (0.190)
SVD_121	-	-	-0.484^{**} (0.239)	-0.087^{***} (0.012)
SVD_122	-	-	0.923^{***} (0.305)	-0.022^{**} (0.010)
SVD_125	-	-	-0.064^{***} (0.022)	-0.842^{**} (0.429)
SVD_131	-	-	-0.424^{*} (0.252)	-
SVD_128	-	-	-	1.036^{***} (0.329)
SVD_129	-	-	-	0.573^{***} (0.220)
SVD_132	-	-	-0.110^{**} (0.055)	-0.822^{*} (0.427)
SVD_135	-	-	-0.760 (0.521)	-
SVD_136	-	-	-	1.523^{**} (0.730)
SVD_137	-	-	0.239^{**} (0.110)	-0.911^{**} (0.370)
SVD_149	-	-	-1.157^{**} (0.480)	-
SVD_150	-	-	0.826^{**} (0.386)	-
SVD_138	-	-	-	0.278 (0.183)
SVD_143	-	-	-	-0.702 (0.444)
SVD_144	-	-	-	0.951^{*} (0.488)
SVD_145	-	-	-	0.907 (0.564)

SVD_152	-	-	-	0.091 (0.056)
SVD_153	-	-	-	0.593 (0.407)
SVD_159	-	-	-	−0.913** (0.385)
SVD_162	-	-	-	0.785*** (0.286)
SVD_168	-	-	-	1.802*** (0.592)
SVD_170	-	-	-	−0.127** (0.060)
SVD_172	-	-	-	−0.087** (0.040)
SVD_175	-	-	-	−1.136 (0.717)
SVD_178	-	-	-	0.040*** (0.011)
SVD_180	-	-	-	0.390 (0.249)
SVD_188	-	-	-	−0.592* (0.324)
SVD_189	-	-	-	0.745 (0.479)
SVD_191	-	-	-	−0.020 (0.013)
SVD_193	-	-	-	1.143* (0.654)
SVD_194	-	-	-	−0.172 (0.113)
Log Likelihood	−38,767.130	−38,719.340	−38,656.690	−38,592.120

Nota: ± Significance: *p<0.1; Significance: **p<0.05; Significance: ***p<0.01