

Avaliação de características que influenciam nos votos de utilidade de opiniões sobre serviços em Português

Alternative Title: Assessment of features influencing the voting for opinions' helpfulness about services in Portuguese

Augusto C. S. Martins
Programa de Pós-Graduação em
Computação Aplicada
Universidade Tecnológica
Federal do Paraná - UTFPR
augustomartins@alunos.utfpr.edu.br

Cesar A. Tacla
Programa de Pós-Graduação em
Computação Aplicada
Universidade Tecnológica
Federal do Paraná - UTFPR
tacla@utfpr.edu.br

RESUMO

Este trabalho apresenta a aplicação 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 sobre o produto e/ou serviço, 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 valor e localização do hotel e com notas mais baixas para limpeza e qualidade dos quartos. Textos com valores pequenos para inteligibilidade (mais difíceis) recebem mais votos do que textos com valores grandes de inteligibilidade.

Palavras-Chave

Mineração de opiniões, Utilidade da opinião, Qualidade da opinião

ABSTRACT

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 cleanliness and quality of rooms. Texts with small values for intelligibility (more

difficult) receive more votes than texts with large values of intelligibility.

Categories and Subject Descriptors

H.3.1 [Information Systems]: Information Storage and Retrieval—*Content Analysis and Indexing*; I.1.2 [Computing methodologies]: Artificial intelligence—*Information extraction, Language resources*

General Terms

Experimentation

Keywords

Opinion mining, Opinion helpfulness, Opinion quality

1. INTRODUÇÃO

O grande número de opiniões geradas por usuários em *blogs*, redes sociais, *microblogs*, 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 tão pouco é simples de ser processado automaticamente. No entanto, como o número de opiniões cresce rapidamente, cada vez é maior o esforço do usuário para encontrar a informação desejada [22, 16]. Esse excesso de informação pode sobrecarregar os usuários. Além de numerosas, muitas das opiniões são fraudulentas, não-informativas, incompletas ou repetitivas. Além disso, opiniões úteis estão misturadas com um grande número de opiniões sem utilidade.

Uma alternativa utilizada por algumas empresas (i.e. Amazon.com, TripAdvisor, Booking.com) 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 posicionada na sequência do conteúdo da opinião e somente com as opções "Sim" ou "Não" como resposta. Além disso, 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 [10].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil
Copyright SBC 2015.

A tarefa mais importante quando se organiza um conjunto de opiniões é determinar quais serão úteis para os usuários. A maioria dos *sites* apresenta o conteúdo utilizando uma ordenação por data ou pontuação (i.e. nota ou números de estrelas atribuído pelo público) e também apresentam um resumo da avaliação de todos que votaram, como por exemplo "20 de 30 pessoas acham essa opinião útil" ou simplesmente apresentam o número de votos de utilidade que a opinião recebeu junto com o conteúdo do texto [11, 27]. No geral, as opiniões são criadas por pessoas que possuem necessidades, pontos de vista e tiveram experiências diferentes sobre um mesmo objeto e revelam quais atributos do produto ou serviço são os mais importantes. 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.

Portanto, a contribuição desta pesquisa é apresentar uma aplicação da metodologia descrita no trabalho de [5] em outra língua (português), com opiniões no domínio de serviços (hotelaria) e a utilização de métricas de inteligibilidade, a fim de identificar entre características básicas (ex. nota sobre o produto e/ou serviço, data da publicação), textuais (ex. tamanho das palavras, parágrafos), semântica (ex. o significado das palavras do texto) do conteúdo de opiniões de serviços de hotel, escritas em português, quais possuem maior influência nos usuários em sua avaliação de utilidade, a fim de definir um modelo capaz de selecionar um conjunto de opiniões que possuem maior probabilidade de receber votos.

O restante do artigo está organizado da seguinte maneira: a Seção 2 apresenta os trabalhos correlatos; a Seção 3 apresenta a metodologia, a seção 4, os experimentos realizados e os resultados obtidos; e as considerações finais na Seção 5.

2. TRABALHOS CORRELATOS

A revisão do estado da arte realizada a partir do estudo de [16], resultou na descoberta do trabalho de [5], através de buscas realizadas nas bases de publicações científicas do *IEEE*, *ACM*, *Cite Seer X*, *Web Of Knowledge*, Periódicos Capes, com palavras-chave relativas a mineração de opiniões, análise da qualidade, utilidade da opinião e qualidade da opinião (e equivalentes em inglês¹). A partir dos resultados encontrados, foram selecionados trabalhos relacionados a qualidade da opinião no período de 2006 à 2014 ordenados por número de citações.

Na literatura, é possível encontrar trabalhos [11, 29, 9, 17, 18, 21, 28, 6, 19] que têm por objetivo estimar automaticamente a qualidade das opiniões através do número de votos de utilidade e outros atributos (ex. frequência de palavras, classes gramaticais) capturados por meio de mineração de dados e aprendizado de máquina, e baseadas no conteúdo das opiniões que receberam pelo menos um voto de utilidade.

Também há uma linha de pesquisa [26, 5, 10, 12, 15, 27] que tenta responder a pergunta: "quais os fatores que influenciam no número de votos que uma opinião recebe?". Esses trabalhos também utilizam técnicas de mineração de dados e aprendizado de máquina, porém consideram todas as opiniões e procuram analisar o quanto as propriedades básicas, textuais, erros de escrita, semântica, histórico de contribui-

ções 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 *online*.

Os artigos de [5] e [12] foram considerados pontos de partida para este trabalho, pois, o primeiro utiliza somente o conteúdo disponível de forma pública no *site* CNETD Download.com, sem a necessidade de informações restritas dos autores das opiniões, interação em redes sociais entre usuários e informações mais detalhadas das opiniões (i.e. análise de subjetividade, polaridade, etc.) e o segundo utiliza métricas de inteligibilidade da língua inglesa, a subjetividade das opiniões, os erros de escrita e o histórico de votos de utilidade que o autor da opinião recebeu, com o objetivo de determinar se essas características dos textos têm influência sobre o volume de vendas dos produtos analisados.

O trabalho de [12] afirma que testes de inteligibilidade são utilizados para quantificar diferentes tipos de texto dentro das áreas da ciência da informação. Em sua pesquisa, os autores aplicam quatro testes de inteligibilidade disponíveis para a língua inglesa: a) *Gunning's Fog* (FOG), b) *Flesh/Kincaid Reading* (FK), c) *Automated Readability-Index* (ARI), c) *Coleman-Liau Index* (CLI) e afirmam que as características textuais é que possuem maior influência na avaliação de utilidade das opiniões. Porém, em sua pesquisa [5], chega a conclusão que as características semânticas apresentam maior influência.

Além das conclusões diferentes, os autores utilizaram metodologias diferentes em dois pontos: (i) o primeiro não utiliza a Análise de Semântica Latente (*LSA*, do inglês *Latent Semantic Analysis* [13, 8]) e (ii) o segundo não utilizou métricas de inteligibilidade. Essa divergência cria uma lacuna para a investigação efetiva de quais características do texto são consideradas pelos usuários quando lêem e avaliam uma opinião como útil ou não. Contudo, as métricas de inteligibilidade da língua inglesa não se aplicam a língua portuguesa [4, 25]. Surge daí a pergunta: "quais os fatores que influenciam no número de votos que uma opinião escrita em português recebe?". Essa é a pergunta na qual se baseia este trabalho.

3. MÉTODO ADOTADO

Os experimentos foram realizados seguindo a sequência utilizada no trabalho de [5], porém adaptada para opiniões escritas em português e do domínio de serviços (hotéis). Os passos realizados foram: pré-processamento do texto para separar as características e realizar contagens textuais (ex. números de palavras, número de sentenças), a análise sintática, redução de termos através da *LSA* e a Análise de Fatores.

Dessa forma, este trabalho estende a pesquisa de [5] de três maneiras: (i) avaliar se é possível chegar a resultados aproximados, porém, utilizando opiniões no domínio de serviços de hotéis no lugar de *software* de computador (i.e. produtos, domínio utilizado em [5]); (ii) aplicar métricas de inteligibilidade utilizando uma implementação da ferramenta *Coh-Metrix-Port* de [25], para verificar se o nível de dificuldade de entendimento de textos influencia nos votos de utilidade; e (iii) utilizar a língua portuguesa² como idioma principal na avaliação das características textuais e semânticas das

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

²Esta pesquisa não distingue opiniões em Português do Brasil de outras variações da Língua Portuguesa.

opiniões.

3.1 Métricas de Inteligibilidade

Métricas de inteligibilidade, como as baseadas no número de sílabas por palavra e na média de palavras por sentenças em parágrafos, nem sempre traduzem a complexidade do texto. As técnicas de processamento de linguagem natural (PLN) possuem ferramentas capazes de verificar a inteligibilidade e automatizar a tarefa de classificar um texto de acordo com o nível de escolaridade ou audiência [24]. Existem somente duas ferramentas de avaliação da inteligibilidade para o Português, sendo estas, uma adaptação do índice *Flesh Reading Ease* (Martins et al. (1996) *apud* [25]) e uma nova abordagem de análise da inteligibilidade baseada na simplificação de textos para pessoas com baixo nível de escolaridade [25].

A ferramenta Coh-Metrix (Graesser et al. (2004), McNamara et al. (2002), Crossley et al. (2007) *apud* [25]), 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. (2003) *apud* [25] como “*características de um texto que, de alguma forma, ajudam o leitor a conectar mentalmente as idéias do texto*”. Já coerência é definida pelo mesmo autor como “*características do texto (ou seja, aspectos de coesão) que provavelmente contribuem para a coerência da representação mental*”.

Devido à incompatibilidade das métricas de inteligibilidade da língua inglesa com a língua portuguesa, foi necessário utilizar um conjunto de métricas adaptadas para o português [4], o Coh-Metrix-Port, desenvolvido por [25].

3.2 Coleta de dados

Os dados para esta pesquisa foram coletados de opiniões reais dos serviços de hotéis do *site* de viagens TripAdvisor, que indexa entre outras informações, as opiniões dos viajantes de diversas cidades em mais de 40 países em todo o mundo. As opiniões são exibidas em uma área pública do *site* que também possui informações do hotel (i.e. endereço, fotos, preços, avaliação geral), dispostas na forma de uma lista (ordenada da mais recente para a mais antiga) e limitada a 10 opiniões por página.

Aproveitando a escolha do Brasil para realização da Copa do Mundo de 2014, foi coletada e utilizada uma amostra aleatória com 10.920 opiniões de 2.457 hotéis das 12 cidades sede selecionadas para o evento do futebol: 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 [5] foram utilizadas 3.460 opiniões e em [12], 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 1.

Conforme discutido no trabalho de [5], a maioria das opiniões não recebe voto de utilidade. O problema, que se repete em outros trabalhos citados anteriormente, também foi encontrado na distribuição de votos por opinião neste trabalho (Tabela 2). É possível notar que quando a quantidade de votos aumenta, o número de opiniões diminui expressivamente. A quantidade de opiniões que nunca recebeu voto ultrapassa 60% da base, enquanto a quantidade que recebeu

dez ou mais votos, não alcança 1% do total de opiniões.

3.2.1 Opinião dos usuários do TripAdvisor

A Figura 1 é um exemplo de uma opinião registrada no TripAdvisor. No canto superior esquerdo abaixo do título está a avaliação geral do hotel 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, quando o usuário preenche esses dados, estão posicionados antes da pergunta sobre a utilidade da opinião (circulado). O TripAdvisor apresenta somente o número total de votos que a opinião recebeu, não havendo diferença entre votos positivos ou negativos.

3.2.2 Características básicas das opiniões

As características básicas podem ser extraídas de forma simples e sem a utilização de ferramentas de PLN (informações disponíveis na Figura 1). Foram utilizados: 1) o intervalo (em dias) entre a data da publicação da opinião e a data da viagem, 2) o extremismo da avaliação, 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: 3) custo-benefício, 4) localização, 5) qualidade do sono, 6) quartos, 7) limpeza e 8) atendimento.

Tabela 1: Quantidade de opiniões por cidade.

Cidade	Hotéis (Qtd.)	Opiniões (Qtd.)	%
Belo Horizonte	159	608	5.57
Brasília	87	673	6.16
Cuiabá	58	135	1.24
Curitiba	143	823	7.54
Fortaleza	219	903	8.27
Manaus	80	286	2.62
Natal	220	1038	9.51
Porto Alegre	111	536	4.91
Recife	98	552	5.05
Rio de Janeiro	529	2123	19.44
Salvador	246	831	7.61
São Paulo	507	2412	22.09
Total	2457	10920	100

3.2.3 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 este trabalho, foi implementada uma versão da ferramenta Coh-Metrix-Port, baseada na documentação disponível *online*³, somente com as contagens básicas (i.e. cálculo da inteligibilidade, número de sentenças, número de palavras, número de sílabas, média de palavras por sentenças, número de verbos, número de substantivos, número de adjetivos, número de advérbios e número de pronomes).

O índice *Flesh Reading Ease* 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 sendo a única métrica que foi adaptada para o português [4]. A saída da fórmula *Flesh Reading*

³<http://www.nilc.icmc.usp.br:3000/index/info>

Ease é um número entre 0 e 100, e quanto mais alto o índice, mais fácil a leitura [4, 25]:

$$206.835 - (1.015 \times ASL) - (84.6 \times ASW) \quad (1)$$

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:

$$248.835 - (1.015 \times ASL) - (84.6 \times ASW) \quad (2)$$

que corresponde à fórmula original salvo pela constante 248,835 que diferencia textos em inglês de textos em português⁴. De acordo com [25], os valores desse índice variam entre 100-75 (muito fácil), 75-50 (fácil), 50-25 (difícil) e 25-0 (muito difícil), que correspondem, respectivamente, às duas subdivisões da educação fundamental (1-4 e 5-8), ensino médio (9-11) e ensino superior. A partir dessa implementação simplificada do Coh-Metrix-Port, foi possível extrair as características textuais, tais como: 1) índice de inteligibilidade (em uma escala de 0 a 100) e as contagens de 2) sentenças, 3) palavras, 4) sílabas, 5) média de palavras por sentenças, 6) quantidade de verbos, 7) substantivos, 8) adjetivos, 9) advérbios e 10) pronomes.

Tabela 2: Quantidade de opiniões por votos de utilidade.

Votos de Utilidade	Opiniões (Qdt.)	%
0	6808	62.34
1	2479	22.7
2	882	8.08
3	340	3.11
4	154	1.41
5	73	0.67
6	35	0.32
7	42	0.38
8	15	0.14
9	17	0.16
≥ 10	75	0.69
Total	10920	100

3.2.4 Características semânticas das opiniões

As características semânticas representam a ideia principal da opinião e, segundo [5], causam grande influência no voto de utilidade da opinião. Para capturar a ideia principal das opiniões, foi utilizada a LSA, que é um método estatístico que explora a relação existente entre um conjunto de termos e os documentos nos quais estes aparecem, a fim de construir um padrão de relacionamento entre os termos e os documentos [7].

3.3 Pré-processamento das opiniões

Na etapa de pré-processamento, foram eliminadas todas as opiniões fora do intervalo de publicação definido entre 1º

⁴O número 42 é o valor de referência que, na média, diferencia textos em inglês e português citado por Martins et al. (1996) *apud* [25]

“Que lugar maravilhoso, super indico!”

●●●●● Avaliou em Abril 3, 2014

Que lugar maravilhoso gente!! Eu e meu marido fomos muito bem recebidos por todos que trabalham na pousada. Com isso o atendimento foi nota 10. A pousada está impecável, jardins bem cuidados, apartamentos limpos. A área da piscina é fantástica com seus ofurôs e banheiras com vista para as praias de Bombas e Bombinhas. A praia que fica bem perto da pousada é linda com água cristalina e areia branca. Nosso final de semana foi incrível. Indico à todos que se hospedem nesta linda pousada em Bombinhas - SC Abraços

se hospedou em Março 2014, viajou com a família

●●●●● Custo-benefício ●●●●● Quartos
●●●●● Localização ●●●●● Limpeza
●●●●● Qualidade do sono ●●●●● Atendimento

menos

Esta avaliação foi útil? Sim < 11

Figura 1: Exemplo de opinião de uma pousada situada em Bombinhas/SC. Fonte: Captura de tela do site Tripadvisor.

de Janeiro e 1º de Agosto de 2014, com o intuito de abranger um período de tempo razoável antes e durante a Copa do Mundo de Futebol 2014 (realizada entre 12 de Junho e 13 de Julho/2014) e selecionar uma amostra aleatória com no máximo 10.920 opiniões⁵. Também foi calculado o intervalo (em dias) entre a data da publicação da opinião e a data da viagem e o nível de “intensidade”, estimado pela diferença entre a nota da opinião dada ao hotel e a nota média de todos os usuários. O processamento foi realizado utilizando a linguagem R [23] 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).

3.4 Análise sintática e redução de termos

Na etapa de análise sintática, foram realizadas as contagens básicas da estrutura do texto (i.e. contagens de sentenças, palavras, sílabas, média de palavras por sentenças, quantidade de verbos, substantivos, adjetivos, advérbios e pronomes) e o cálculo do índice de inteligibilidade. Na sequência, foram feitas a remoção de *stopwords*⁶ e a redução das palavras ao seu radical com o objetivo de agrupar termos semelhantes, reduzindo a quantidade de palavras e a ambiguidade do texto para a próxima etapa. A ferramenta Apache OpenNLP [3] em conjunto com algoritmos próprios foi utilizada para realizar a classificação sintática e identificação de palavras e sentenças (i.e. processamento “*part-of-speech*”).

A filtragem de termos do texto resultou na construção da matriz termos-documentos. Essa matriz, foi construída utilizando a *TF-IDF* (do inglês, *Term Frequency-Inverse Document Frequency*). Esse processo consiste na separação do

⁵Essa quantidade de opiniões é o triplo do número de opiniões que [5] utiliza em seu trabalho.

⁶Palavras que podem ser consideradas irrelevantes. Exemplos: as, e, os, de, para, com, sem, foi, etc

conteúdo dos documentos (i.e. opiniões) em termos e a contagem de suas frequências. O resultado deste processamento gera uma matriz esparsa (conhecida com *Bag of Words* ou *BOW*) em que cada linha representa um documento e cada coluna um termo, isto é, seja d_i a linha e t_j a coluna da matriz, o elemento da matriz O_{ij} representa o número de vezes que o termo j aparece no documento i . Foram identificados 7.976 termos, resultando em uma matriz 10.920x7.976.

3.5 Análise semântica das opiniões

Na sequência da redução de termos e criação da matriz termos-documentos, [5] aplica a Decomposição de Valor Singular [14] (do inglês, *Singular Value Decomposition* ou SVD), sendo esta a etapa principal da LSA. Através desta técnica (SVD) são geradas três matrizes a partir da matriz termos-documentos original, e então uma nova matriz é obtida do produto destas três matrizes. A SVD é utilizada para localizar a informação semântica essencial em uma matriz de co-ocorrência das palavras. O objetivo é obter um espaço semântico condensado que representa a melhor relação entre os termos e documentos.

3.6 Análise Fatorial

A Análise Fatorial (do inglês, *Factor Analysis*) é um processo de mensuração que identifica variáveis similares em uma mesma estrutura. A denominação "Fatores SVD" é utilizada por [5] para definir as colunas da nova matriz de termos encontrada após à análise fatorial. Essa matriz resume ainda mais a matriz-semântica das opiniões e representa o conjunto de características semânticas que serão utilizadas nos modelos, isto é, um número reduzido de palavras que representam o conteúdo textual da opinião e que foram encontradas devido similaridades matemáticas dos termos, representando o nível mais simplificado da matriz termos-documento. A pesquisa de [5], utiliza conjuntos de 50, 100, 150 e 200 Fatores SVD, porém, o resultado apresentado para fins de demonstração utiliza somente 100 fatores, o mesmo número utilizado neste trabalho.

3.7 Definição dos modelos

A Regressão Logística Ordinal [20] [2], do inglês *Ordinal Logistic Regression* ou OLR, é utilizada por [5] para investigar a relação entre as características básicas, textuais e semânticas das opiniões e o número de votos de utilidade da opinião. A OLR é uma extensão do modelo de regressão logística binária, onde a variável dependente pode acomodar mais do que duas categorias de valores. A variável dependente deste estudo é um "ranking de utilidade" baseado no número de votos que uma opinião recebeu. O modelo OLR foi utilizado para determinar quais características implicam no aumento de votos de utilidade de uma opinião. As equações 3 e 4 definem a abordagem da OLR:

$$g(Pr(Y \leq i|x)) = \alpha_i + \beta'x, i = 1, 2, \dots, k \quad (3)$$

onde Y é a variável resposta (dependente) com k categorias codificadas em $1, 2, \dots, k$, $\alpha_1, \alpha_2, \dots, \alpha_k$, representa o parâmetro de localização para as k categorias da variável dependente, β é o vetor de coeficientes de regressão (declives), β' é o vetor β transposto e x é matriz das variáveis independentes. A função $g=g(\mu)$ é a função de ligação que permite (μ) assumir uma resposta linearmente relacionada as variáveis independentes. $Pr(Y \leq i|x)$ é a probabilidade de Y ser

Tabela 3: Descrição dos Modelos

Modelo	Descrição das variáveis	# variáveis
1	Básicas	8
2	Textuais	10
3	Semânticas (Fatores SVD)	100
4	Básicas + Textuais	18
5	Básicas + Textuais + Semânticas	118

menor ou igual a i , dado um valor de x . A escala de *odds* proporcionais foi usada como função de ligação, conforme definido na equação 4 [5]:

$$\begin{aligned} g_i(Pr(Y \leq i|x)) &= \ln \frac{P_r(Y \leq i|x)}{P_r(Y > i|x)} = \ln \frac{P_r \leq i|x}{1 - P_r(Y \leq i|x)} \\ &= \ln \frac{\odot_1(x) + \odot_2(x) + \dots + \odot_i(x)}{1 - (\odot_1(x) + \odot_2(x) + \dots + \odot_i(x))} \\ &= \alpha_i + \beta'x, \\ i &= 1, 2, \dots, k \end{aligned} \quad (4)$$

onde $\odot_i(x)$ é a probabilidade de se estar em uma classe i dado x . Nesta pesquisa, Y é o número de votos de utilidade das opiniões e estes foram agrupados de forma que os valores de "0" a "9" representam estes valores neste intervalo, enquanto "10" agrega "10 ou mais votos" em uma mesma categoria. A distribuição dos votos foi apresentada anteriormente na Tabela 2.

Para determinar como e quais características de uma opinião influenciam no número de votos de utilidade, foram criados cinco modelos OLR com várias combinações entre os três tipos de características básicas, textuais e semânticas seguindo a mesma combinação discutida no trabalho de [5]. A Tabela 3 apresenta a descrição dos modelos e a quantidade de variáveis independentes de cada um. Os três primeiros modelos (i.e. modelos 1, 2 e 3) serão utilizados para verificar a influência de cada característica separadamente. O modelo 4 congrega características básicas e textuais e o modelo 5 engloba os três tipos de características.

Alguns dos modelos possuem uma grande quantidade de variáveis, o que pode aumentar o ajuste do modelo e diminuir sua capacidade de previsão. Por isso é utilizado o método de seleção *stepwise*, que retira do modelo uma variável com maior *p-valor* e verifica-se se houve diminuição no ajuste do modelo. Caso não tenha ocorrido efeito no ajuste, a variável é retornada para o modelo e o mesmo processo é executado para a próxima variável. Esse processo é executado até que todas as variáveis tenham sido testadas ou que o modelo atinja seu ajuste mínimo. Para comparar os modelos, [5] utilizou três medidas de ajuste: taxa de classificação incorreta (do inglês, *misclassification rate*), o critério de informação de Akaike (i.e. *Akaike's information criterion*, ou AIC) e a razão de lift (i.e. *lift ratio*).

A taxa de classificação incorreta, é utilizada para calcular o quanto uma classificação pode ser considerada incorreta. É calculada 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. Criado por Akaike [1], o critério de informação é utilizado para manter o compromisso entre a precisão do modelo e sua complexidade. Seu cálculo penaliza a inclu-

Tabela 4: Comparação dos Modelos

Modelo	Taxa de Classificação Incorreta	AIC	Lift ratio
1	0,62344	24651,57	0,98
2	0,37595	24602,58	1,18
3	0,37546	24712,84	0,88
4	0,37568	24588,56	1,16
5	0,37445	24530,49	1,27

são de muitas variáveis no modelo enquanto procura manter uma precisão razoável de maneira que, quanto menor o índice AIC, melhor o modelo. Esta métrica foi utilizada para comparar os cinco modelos criados para esta pesquisa. A razão de lift é utilizada para medir a performance de previsão de um modelo em um segmento aleatório da população. Como regra geral, quanto maior, melhor será o desempenho do modelo.

4. RESULTADOS OBTIDOS

Nesse estudo, chegou-se a uma conclusão semelhante a de [5] em que as características semânticas, ou seja, grupos de palavras encontradas no texto da opinião, possuem influências positivas e negativas no número de votos de utilidade que recebem. Complementar a essa conclusão, foi verificado que as pessoas tendem a dar mais atenção para opiniões mais recentes, que tenham notas mais altas em relação ao valor, localização do hotel e que apresentem notas mais baixas em relação à limpeza e qualidade dos quartos. Curiosamente, textos de opiniões com mais palavras, menos adjetivos e índices de inteligibilidade mais baixos (i.e. maior dificuldade) recebem mais votos do que opiniões sem estas características.

A Tabela 4 resume os resultados empíricos da comparação dos cinco modelos OLR. Assim como no trabalho de [5], o Modelo 5 (i.e. Básicas + Textuais + Fatores SVD), possui a menor taxa de classificação incorreta e os melhores índices AIC e *lift*, ou seja, obteve o melhor resultado entre os modelos. Esse resultado confirma que integrar as características semânticas junto com as características básicas e textuais, melhora o desempenho do modelo.

O Modelo 4 (i.e. Básicas + Textuais) possui o segundo melhor AIC, mas fica com a terceira posição em duas outras métricas (taxa de classificação incorreta e *lift*). O Modelo 2 (Textuais), com o terceiro melhor AIC, alterna entre a segunda e terceira posição para os índices *lift* e taxa de classificação incorreta. Os Modelos 1 (Básicas) e 3 (Fatores SVD), possuem os piores resultados em duas métricas (AIC e *lift*), o que pode indicar que características básicas e semânticas utilizadas isoladamente são os piores critérios para estimular o voto de utilidade.

A Tabela 5, mostra o resultado final do Modelo 5 com as variáveis selecionadas após a aplicação do método de seleção de variáveis da OLR, suas estimativas, desvio padrão, teste estatístico *Wald Chi-square* e o *p-valor*. Este modelo possui cinco variáveis básicas (intervalo, valor, quartos, limpeza, localização), cinco variáveis textuais (índice *Flesch Reading Ease*, número de palavras, número de substantivos, número de adjetivos e número de advérbios) e 47 variáveis semânticas (fatores SVD).

É importante notar que o *p-valor* de mais da metade das variáveis é menor que 0,05, o que indica que são estatisticamente significativas com 95% de nível de confiança. Porém, uma variável básica (número de palavras) e quatro variá-

veis semânticas⁷ tiveram o *p-valor* menor que 0,01, o que demonstra que são estatisticamente significativas com 99% de nível de confiança. A maioria dos *p-valores* do Modelo 5 está no intervalo de 0,05 e 0,01 de significância estatística com níveis de confiança entre 95% e 98% respectivamente.

Tabela 5: Estimativa dos parâmetros do Modelo OLR 5 (Básicas + Estilo + SVD)

Parâmetro	Coef.	Erro Padrão	Wald Chi-square	Pr > ChiSq
Intervalo	-0,0004	0,0002	6,72	0,0095
Valor	0,0512	0,0209	6,00	0,0143
Quartos	-0,0419	0,0223	3,53	0,0603
Limpeza	-0,0607	0,0236	6,59	0,0103
Localização	0,0305	0,0197	2,40	0,1210
Flesch R.E.	-0,0022	0,0010	5,33	0,0209
N. Palavras	0,0112	0,0024	22,78	< 0,0001
N. Substantivos	-0,0156	0,0072	4,72	0,0298
N. Adjetivos	-0,0164	0,0079	4,32	0,0376
N. Advérbios	-0,0213	0,0099	4,60	0,0320
SVD_4	0,1002	0,0405	6,12	0,0134
SVD_6	0,1356	0,0714	3,61	0,0575
SVD_9	0,4612	0,1864	6,12	0,0133
SVD_12	-0,3343	0,1294	6,68	0,0098
SVD_20	-0,2316	0,1553	2,23	0,1358
SVD_24	0,4922	0,3125	2,48	0,1152
SVD_21	0,7037	0,2110	11,12	0,0009
SVD_25	0,4151	0,2541	2,67	0,1024
SVD_33	0,0871	0,0592	2,16	0,1413
SVD_34	-0,4320	0,2860	2,28	0,1309
SVD_32	0,1636	0,0465	12,35	0,0004
SVD_35	-0,4477	0,2823	2,52	0,1127
SVD_36	0,6797	0,3603	3,56	0,0592
SVD_44	0,1556	0,0793	3,85	0,0498
SVD_43	0,6344	0,3186	3,96	0,0465
SVD_57	-0,0425	0,0238	3,19	0,0742
SVD_50	1,1196	0,5226	4,59	0,0322
SVD_77	-0,0878	0,0389	5,10	0,0240
SVD_58	0,2821	0,1207	5,46	0,0194
SVD_99	-0,0565	0,0306	3,40	0,0651
SVD_82	-0,3221	0,2178	2,19	0,1392
SVD_69	-0,0334	0,0188	3,15	0,0760
SVD_84	-0,0533	0,0254	4,41	0,0357
SVD_86	-0,3533	0,1852	3,64	0,0565
SVD_67	0,7409	0,2480	8,92	0,0028
SVD_83	-0,9946	0,5425	3,36	0,0668
SVD_100	1,0611	0,4500	5,56	0,0184

Ao contrário do trabalho de [5], 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 palavras e negativo para as outras variáveis textuais, ou seja, uma grande quantidade de palavras e um menor índice de inteligibilidade (i.e. mais difícil) com poucos substantivos, adjetivos e advérbios possui maior influência no número de 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.

As opiniões "extremamente negativas" podem não exercer o mesmo efeito no domínio de serviços ou em outra cultura. Ao contrário do resultado encontrado em [5], que observou que as pessoas tendem a votar nas opiniões mais negativas, o Modelo 5 deste trabalho não incluiu a variável relacionada ao extremismo da opinião. A avaliação geral da nota pode ter dado lugar a avaliações das características específicas do hotel, conforme explicado a seguir. As variáveis contendo as notas sobre "valor", "quarto", "limpeza" e "localização" são

⁷SVD_12: "abertur"; SVD_21: "abram"; SVD_32: "abrind"; SVD_67: "aceit".

importantes para quantidade de votos de utilidade da opinião, ou seja, são as variáveis básicas que ajudam a diminuir o índice AIC sem prejudicar o desempenho do modelo. Por possuir uma estimativa positiva, os usuários tendem a votar em opiniões com notas mais altas para "valor" e "localização".

Apresentando estimativas negativas, valores mais baixos para "limpeza" e "quarto" atraem mais votos, o que indica tendência de procura por serviços com conforto e higiene. Estas quatro variáveis, podem indicar as características dos serviços de hotel que os usuários levam em consideração na escolha da empresa que irão contratar e estão diretamente ligadas ao domínio utilizado, isto é, certamente a utilização de opiniões sobre restaurantes terá outras variáveis básicas. Finalmente, os coeficientes de alguns fatores SVD (Seção 3.6) são positivos e outros negativos, o que indica que certas palavras possuem influência positiva, atraindo mais votos e outras palavras possuem o efeito contrário, desestimulando que o usuário avalie a opinião.

5. CONCLUSÃO

Neste trabalho foram realizados os experimentos descritos na pesquisa de [5] para encontrar quais características em um conjunto de opiniões pode influenciar a percepção de utilidade de uma opinião e, conseqüentemente, na quantidade de votos de utilidade que uma opinião pode receber. Assim como no trabalho de [5], o objetivo foi explorar se as características semânticas possuem influência no número de votos de utilidade. Foi possível observar nos resultados empíricos da pesquisa que essas as características semânticas efetivamente possuem impacto na avaliação de utilidade das opiniões também em língua portuguesa e no domínio de serviços da hotelaria. Curiosamente, a utilização do índice de inteligibilidade revelou que opiniões mais complexas e que possuem mais palavras são mais úteis do que opiniões com poucas palavras e mais fáceis de ler.

Estendendo o experimento de [5] e considerando algumas de suas limitações, foi possível adaptá-lo à mudança da Língua Inglesa para a Portuguesa e aplicá-lo em outro domínio (hotelaria). Assim como mencionado na conclusão de [5], as características semânticas são importantes para a percepção de utilidade das opiniões, mas quais são efetivamente as palavras que capturam a atenção dos leitores? Essa investigação pode levar a descoberta de padrões de textos que potencializam as chances de leitura e mais votos em uma determinada opinião.

As opiniões utilizadas são específicas para hotéis e possuem suas características próprias, como localização do hotel, detalhes referentes aos quartos, atendimento dos funcionários e outros, porém, a base de dados também possui informações sobre opiniões de Restaurantes e Pontos Turísticos das cidades escolhidas para a Copa do Mundo de 2014 e sendo possível comparar em um trabalho futuro se os resultados se mantêm, por exemplo, para opiniões sobre atrações turísticas. Além disso pode-se investigar a existência de um índice de inteligibilidade "recomendado" e que teria um efeito semelhante na quantidade de votos de utilidade. Outro trabalho futuro, seria retirar a limitação arbitrária de 100 Fatores SVD (utilizada no artigo de [5]) e investigar qual o conjunto mínimo de fatores que melhoram o desempenho do modelo penalizá-lo pela inclusão de um número excessivo de variáveis.

6. AGRADECIMENTOS

Os autores agradecem à Fundação Araucária, CAPES e CNPq pelo financiamento parcial deste trabalho.

7. REFERÊNCIAS

- [1] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, Dec 1974.
- [2] J. A. Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–30, 1984.
- [3] Apache. Opennlp, 2011.
- [4] E. M. F. Barboza and E. M. d. A. Nunes. A inteligibilidade dos websites governamentais brasileiros eo acesso para usuários com baixo nível de escolaridade. *Inclusão Social*, 2(2):19–33, 2007.
- [5] Q. Cao, W. Duan, and Q. Gan. Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, 50(2):511–521, Jan. 2011.
- [6] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web*, page 141–150. ACM, 2009.
- [7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41(6):391–407, 1990.
- [8] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *JASIS*, 41(6):391–407, 1990.
- [9] A. Ghose and P. G. Ipeirotis. Designing novel review ranking systems: predicting the usefulness and impact of reviews. In *Proceedings of the ninth international conference on Electronic commerce*, page 303–310. ACM, 2007.
- [10] A. Ghose and P. G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering*, 23(10):1498–1512, Oct. 2011.
- [11] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, page 423–430. Association for Computational Linguistics, 2006.
- [12] N. Korfiatis, E. García-Bariocanal, and S. Sánchez-Alonso. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications*, 11(3):205–217, May 2012.
- [13] T. K. Landauer, P. W. Foltz, and D. Laham. *Introduction to Latent Semantic Analysis*, volume 25, pages 259–284. Discourse Processes, 1998.
- [14] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284, 1998.

- [15] J. Lee. What makes people read an online review? the relative effects of posting time and helpfulness on review readership. *Cyberpsychology, Behavior, and Social Networking*, 16(7):529–535, July 2013.
- [16] B. Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.
- [17] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, page 334–342, 2007.
- [18] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. pages 443–452. IEEE, Dec. 2008.
- [19] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, page 691–700. ACM, 2010.
- [20] P. McCullagh. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, pages 109–142, 1980.
- [21] M. P. O’Mahony and B. Smyth. Learning to recommend helpful hotel reviews. In *Proceedings of the third ACM conference on Recommender systems*, page 305–308. ACM, 2009.
- [22] B. Pang and L. Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, page 115–124. Association for Computational Linguistics, 2005.
- [23] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. ISBN 3-900051-07-0.
- [24] C. Scarton, C. Gasperin, and S. Aluisio. Revisiting the readability assessment of texts in portuguese. In *Advances in Artificial Intelligence–IBERAMIA 2010*, page 306–315. Springer, 2010.
- [25] C. E. Scarton and S. M. Aluísio. Análise da inteligibilidade de textos via ferramentas de processamento de língua natural: adaptando as métricas do coh-metrix para o português. *Linguamática*, 2(1):45–61, 2010.
- [26] A. Talwar, R. Jurca, and B. Faltings. Understanding user behavior in online feedback reporting. In *Proceedings of the 8th ACM conference on Electronic commerce*, page 134–142. ACM, 2007.
- [27] J. Tang, H. Gao, X. Hu, and H. Liu. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM conference on Recommender systems*, page 1–8. ACM, 2013.
- [28] O. Tsur and A. Rappoport. RevRank: a fully unsupervised algorithm for selecting the most helpful book reviews. In *ICWSM*, 2009.
- [29] Z. Zhang and B. Varadarajan. Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, page 51–57. ACM, 2006.