UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

MATHEUS ERNESTO SILVA GONÇALVES
THAIS LULLEZ

DETECÇÃO DE AVALIAÇÕES FALSAS DE HOTÉIS ONLINE

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2018

MATHEUS ERNESTO SILVA GONÇALVES
THAIS LULLEZ

DETECÇÃO DE AVALIAÇÕES FALSAS DE HOTÉIS ONLINE

Trabalho de Conclusão do Curso de Bacharelado em Sistemas de Informação, apresentado à UTFPR como requisito parcial para obtenção do título de bacharel em Sistemas de Informação.

Orientador: Alexandre Reis Graeml

CURITIBA
2018

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Câmpus Curitiba
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Informática
Coordenação do Curso de Bacharelado em Sistemas de Informação

TERMO DE APROVAÇÃO

“DETECÇÃO DE AVALIAÇÕES FALSAS DE HOTÉIS ONLINE”

por

Matheus Ernesto Silva Gonçalves

Thais Lullez

Este Trabalho de Conclusão de Curso foi apresentado como requisito parcial à obtenção do grau de Bacharel em Sistemas de Informação na Universidade Tecnológica Federal do Paraná - UTFPR - Câmpus Curitiba. O(a)(s) aluno(a)(s) foi(ram) arguido(a)(s) pelos membros da Banca de Avaliação abaixo assinados. Após deliberação a Banca de Avaliação considerou o trabalho _____.

_____ **<Prof. Alexandre Reis Graeml >** (Presidente - UTFPR/Curitiba)	_____ **<Prof. Rita Cristina Galarraga Berardi>** (Avaliador(a) 1 - UTFPR/Curitiba)
_____ <**Prof. Thiago Henrique Silva >** (Avaliador 2(a) - UTFPR/Curitiba)	_____ **<Profa. Leyza Baldo Dorini>** (Professora Responsável pelo TCC – UTFPR/Curitiba)
_____ **<Prof. Leonelo Dell Anhol Almeida>** (Coordenadordo curso de Bacharelado em Sistemas de Informação – UTFPR/Curitiba)	

“A Folha de Aprovação assinada encontra-se na Coordenação do Curso.”

RESUMO

É cada vez mais comum que as pessoas leiam as experiências de outros hóspedes de hotéis na Internet e também escrevam e compartilhem suas próprias opiniões. Contudo, nem todos os comentários são verdadeiros. Uma vez que as avaliações positivas podem assegurar fama e significativos ganhos financeiros para os hotéis, assim como avaliações negativas podem denegrir a sua imagem, há algum estímulo para que agentes inescrupulosos se envolvam em práticas fraudulentas de e-business, como o spam de opinião. Este artigo tem como principal objetivo desenvolver um checklist capaz de melhorar a capacidade de detecção de avaliações falsas em sites de compartilhamento de opinião sobre hotéis, tentando garantir a autenticidade das opiniões compartilhadas nessas plataformas.

Palavras-Chave: Spam de opinião, Detecção de spam de opinião, Avaliações de hotéis.

ABSTRACT

It’s becoming more common for people to read as experiences of other hotel users on the Internet and also write and share their own opinions. However, not all reviews are true. Since positive opinions can ensure fame and significant financial gains for hotels, as well as negative opinions can tarnish their image, there is some incentive for unscrupulous agents to engage in fraudulent e-business practices such as opinion spam. This article has as main objective to develop a checklist capable of improving the ability to detect false evaluations in sites of opinion sharing on hotels, trying to guarantee the authenticity of the opinions shared on these platforms.

Keywords: Opinion spam, Opinion spam detection, Hotel reviews.

LISTA DE FIGURAS

LISTA DE TABELAS E QUADROS

LISTA DE ABREVIATURAS E SIGLAS

CNF: Fator de Confiança.

SUMÁRIO

1 INTRODUÇÃO

É cada vez mais comum que as pessoas leiam sobre as experiências de outros consumidores na Internet e também escrevam e compartilhem suas próprias opiniões (Yoo & Gretzel, 2009). No passado, um cliente satisfeito ou insatisfeito costumava comentar sobre sua experiência apenas com poucas pessoas de seu convívio. Agora, graças à popularização de *sites* de *e-commerce* que estimulam seus clientes a avaliar os produtos adquiridos e de sites de compartilhamento de opinião, um único comentário pode influenciar milhares de pessoas. Ao facilitar a difusão dos pontos de vista dos consumidores e o acesso a tais opiniões, a Internet mudou completamente a forma de as pessoas fazerem compras (Bambauer-Sachse & Mangold, 2013; O'Connor, 2008).

Essas opiniões geradas pelos consumidores são especialmente importantes para o setor hoteleiro e de turismo, em geral, cujos produtos intangíveis são difíceis de avaliar antes de seu consumo. Os planos de viagem dos usuários, envolvendo onde ficar e o que fazer, são em grande parte moldados pelas experiências coletivas e opiniões de outros (Litvin et al., 2008; Chua & Banerjee, 2013). O fato de existir um grande número de comentários positivos, com elogios para um hotel, cria uma impressão favorável a respeito do estabelecimento e da experiência que outros tiveram ao frequentá-lo, que pode impulsionar os leitores de um site na Internet, em busca de opções de hospedagem, a fazerem uma reserva. Por outro lado, a existência de comentários negativos pode resultar em uma má impressão, fazendo os possíveis hóspedes escolherem outro lugar para se hospedar (O'Connor, 2008).

Assim, já não são mais as fotografias profissionais e textos sofisticados incluídos em catálogos ou anúncios que definem a decisão de compra dos clientes em potencial do setor hoteleiro, mas comentários e fotos simples postados por hóspedes anteriores em *sites* como o TripAdvisor.com ou o Expedia.com (Bambauer-Sachse & Mangold, 2013; O'Connor, 2008).

Contudo, nem todos os comentários são verdadeiros. Uma vez que as avaliações positivas podem assegurar fama e significativos ganhos financeiros para os hotéis, assim como avaliações negativas podem denegrir a sua imagem, há algum estímulo para que agentes inescrupulosos se envolvam em práticas fraudulentas de *e-business*, como o *spam* de opinião (Banerjee & Chua, 2014). Esse tipo de *spam* envolve a publicação de dois tipos de comentários enganosos: o *spam* de propaganda e o de difamação. O primeiro se refere a opiniões positivas

falsas, que visam a promover um hotel e, consequentemente, aumentar as suas vendas. Já o segundo envolve opiniões negativas falsas, feitas para prejudicar os concorrentes (Jindal & Liu, 2008). Se não forem controladas, tais ações comprometem a qualidade e a utilidade de todo o sistema de avaliação por outros usuários, gerando perdas tanto para empresas que mantêm os serviços de avaliação (compartilhamento de opinião), como para empresas honestas, que não se envolvem em ações antiéticas, e para os próprios consumidores (O'Connor, 2008).

1.1 OBJETIVO GERAL E OBJETIVOS ESPECÍFICOS

Este trabalho tem como principal objetivo desenvolver um checklist capaz de melhorar a capacidade de detecção de avaliações falsas em *sites* de compartilhamento de opinião sobre hotéis, tentando garantir a autenticidade das opiniões compartilhadas nessas plataformas.

Para o atingimento do objetivo geral definido, há que se considerar os seguintes objetivos específicos:

- identificar características de avaliações falsas já analisadas por outros pesquisadores, que possam contribuir na definição de procedimentos para a sua detecção;
- investigar formas de detecção de avaliações falsas adotadas por *sites* de compartilhamento de opinião sobre hotéis;
- projetar um modelo que possibilite a detecção de opiniões falsas capaz de se beneficiar de um checklist;
- criar a base de conhecimento do checklist com base no modelo projetado;
- desenvolver um checklist para detecção de avaliações falsas, apoiado pela base de conhecimentos;
- avaliar o grau de precisão e a vulnerabilidade a falhas do checklist proposto.

1.2 JUSTIFICATIVA

Spam de opinião é um problema generalizado e prejudicial, já que compromete a qualidade e a utilidade de plataformas de compartilhamento de opinião. No setor de turismo, hotéis podem ser prejudicados por críticas negativas falsas. Um viajante, procurando um hotel para se hospedar também pode ser enganado por avaliações falsas (Crawford *et al*., 2015; O'Connor, 2008). Além

disso, a possibilidade de haver opiniões viesadas em *sites* de compartilhamento de opinião pode levar os consumidores a desconfiar das avaliações ali contidas. Isso, por sua vez, pode fazer com que desconsiderem ou subestimem informações que lhes seriam úteis, disponibilizadas por outros viajantes, honestamente interessadas em compartilhar suas experiências (Mayzlin *et al*., 2014).

Portanto, desenvolver métodos para garantir que as opiniões compartilhadas na *web* sejam uma fonte confiável de informação é um trabalho importante, mas desafiador, já que não é possível verificar diretamente se uma avaliação é falsa ou verdadeira (Crawford *et al*., 2015; O'Connor, 2008).

1.3 ESTRUTURA DO TRABALHO

O presente trabalho encontra-se dividido da seguinte maneira, a fim de descrever cada etapa de desenvolvimento da pesquisa necessária para atingir os objetivos propostos:

- Capítulo 2: levanta o estado da arte referente ao tema de estudo, sendo particionado em Avaliações de Hotéis *Online* (Avaliações falsas, Detecção de avaliações falsas e características de Avaliações falsas) e Sistemas Especialistas.
- Capítulo 3: apresenta a metodologia empregada na pesquisa, baseada na revisão sistemática da literatura e a identificação das características que podem indicar spam de propaganda e spam de difamação.
- Capítulo 4: descreve a implementação e desenvolvimento do checklist.
- Capítulo 5: expõe a validação do checklist criado, apresentando como foi feita a criação da base de dados para os testes, por meio de um formulário, as limitações que os testes tiveram, análise dos resultados obtidos e o aprimoramento do checklist.
- Capítulo 6: apresenta as considerações finais a respeito do trabalho, suas limitações e possíveis trabalhos futuros.
- Ao final, são apresentadas as referências e dois apêndices.

2 LEVANTAMENTO BIBLIOGRÁFICO E ESTADO DA ARTE

Este capítulo apresenta o levantamento bibliográfico realizado para subdiar a pesquisa, sendo dividido nas seguintes seções: 2.1 Avaliações de hotéis *online* (subdividida nas seções: 2.1.1 Avaliações falsas, 2.1.2 Detecção de avaliações falsas e 2.1.3 Características de avaliações falsas) e 2.2 Sistemas especialistas.

2.1 AVALIAÇÕES DE HOTÉIS *ONLINE*

Ao planejar uma viagem, o viajante pode utilizar diversas estratégias para escolher em qual hotel vai se hospedar, como perguntar a amigos, consultar uma agência de viagens ou realizar uma pesquisa na Internet. O que todas essas estratégias têm em comum é o fato que as pessoas frequentemente procuram o conselho de outros para sua tomada de decisão (Sparks & Browning, 2011).

O *marketing* boca-a-boca, tanto positivo quanto negativo, tem o poder de influenciar na decisão de compra dos clientes. Como resultado da disseminação da Internet, e da facilidade de produzir conteúdo *online*, surgiu uma nova forma de comunicação boca-a-boca: as mídias sociais. Esse novo canal de comunicação permite distribuir informação por meio de *sites* de busca, *blogs* ou *sites* específicos de opinião sobre produtos. Além disso, está sendo utilizado de várias formas para o planejamento de viagens (Sparks & Browning, 2011).

Sites como o TripAdvisor, que permitem aos usuários compartilhar suas experiências de viagem por meio de *reviews*, adquiriram imensa popularidade com o passar dos anos (Lo *et al*., 2011). As avaliações que proporcionam são bastante influentes, porque são escritas a partir da perspectiva de consumidores, proporcionando uma experiência indireta ao usuário que procura por hospedagem (Yoo & Gretzel, 2009). Os leitores também percebem que outros viajantes estão mais propensos a fornecer informações atualizadas e sinceras, mais confiáveis que as informações fornecidas pelos próprios prestadores de serviços de viagens (Gretzel & Yoo, 2008), que são partes diretamente interessadas no negócio com o potencial cliente. Entretanto, alguns desses *reviews* de viajantes podem ter sido gerados maliciosamente, o que torna um desafio detectar e distinguir avaliações falsas das verdadeiras (Zhou & Sung, 2008).

2.1.1 Avaliações falsas

Avaliações *online* são cada vez mais utilizadas na tomada de decisões de compra. Opiniões positivas podem render ganhos financeiros significativos e fama para as empresas (Banerjee & Chua, 2014). Infelizmente, isso incentiva gente

inescrupulosa a escrever críticas falsas para promover ou desacreditar um determinado produto/serviço. Tais indivíduos são chamados de *spammers* de opinião e sua atividade é chamada de *spam* de opinião ou *review spam* (Jindal & Liu, 2008).

As avaliações falsas geralmente têm os seguintes objetivos (Jindal & Liu, 2008):

- *spam* de propaganda: opiniões positivas não merecedoras de crédito para promover um produto/serviço. No caso de avaliações de hotéis, elas são feitas pelo próprio hotel ou com o apoio/incentivo dele, para promovê-lo positivamente.
- *spam* de difamação: opiniões negativas injustas, maliciosas ou falsas para prejudicar um produto/serviço. No caso de avaliações de hotéis, elas são feitas por concorrentes inescrupulosos com a intenção de prejudicar a reputação de outra empresa no mercado.

A detecção de *spam* é importante para assegurar que *sites* de compartilhamento de opinião continuem a ser fontes confiáveis de opiniões, ao invés de repositórios de informações falsas ou pouco confiáveis (Crawford *et al*., 2015).

2.1.2 Detecção de avaliações falsas

Humanos são menos capazes de detectar mentiras visíveis do que mentiras audíveis (Bond & DePaulo, 2006). A detecção de informações falsas em ambientes *online*, portanto, é mais desafiadora do que em outros ambientes (Zhou & Sung, 2008).

O principal desafio em identificar empiricamente fraudes em *reviews* é que não se pode observar diretamente se uma revisão é falsa. A situação é ainda mais complicada pela falta de um padrão único assumido por *reviews* falsos (Luca & Zervas, 2016). A maioria dos *sites* que contêm avaliações também não tem restrições específicas na publicação de *reviews* e requer poucas informações adicionais dos avaliadores (Schindler & Bickart, 2005). Neste cenário, utilizar tecnologia para ajudar a encontrar detectar avaliações falsas pode representar uma alternativa (Yoo & Gretzel, 2009), desde que haja fatores que proporcionem indícios de fraude em avaliações de hotéis *online* (Yoo & Gretzel, 2009), os quais possam ser identificados com o apoio da tecnologia.

2.1.3 Características de avaliações falsas

Vários estudos têm tentado investigar as diferenças entre avaliações falsas e verdadeiras em ambiente *online* com o propósito de encontrar uma forma de detectar fraudes. Destacam-se os estudos realizados por Yoo & Gretzel (2009), Mukherjee *et al*. (2013), Keates (2007), Filieri (2016), Wu *et al.* (2010), Crawford (2015), Lu *et al*. (2013) e Mukherjee *et al.* (2016).

Em termos de características utilizadas para a detecção de fraudes, existem as características linguísticas e as comportamentais. Diferenças na estrutura de linguagem do texto das avaliações representam o que se tem chamado de características linguísticas, enquanto aspectos do comportamento do *spammer* nas plataformas de compartilhamento de opinião são o que se chama de características comportamentais (Yoo & Gretzel, 2009; Mukherjee *et al*., 2013).

Em relação às características linguísticas, Keates (2007) identifica vários fatores que podem indicar que uma avaliação positiva é falsa, como escores que diferem muito da classificação média e menção a hotéis próximos ao local onde está o usuário informante, como sendo superiores as quais dificilmente teriam sido usados por alguém que mora na região. Do mesmo modo, Mukherjee *et al*. (2013) verificaram que avaliações negativas falsas tendem a ser curtas, porque o *spammer* não tem muito sobre o que escrever. Por esse mesmo motivo, elas também podem ser semelhantes a avaliações anteriores. Os autores também alertam que avaliações extremas, que só expressam sentimentos negativos ou positivos devem ser consideradas suspeitas. Já Filieri (2016) salienta que avaliações que fornecem detalhes que você nunca encontra em uma revisão regular, como o nome do proprietário ou do gerente, são também dignas de desconfiança, da mesma forma que avaliações superficiais, com detalhes irrelevantes sobre a experiência ou com uso abundante de superlativos.

Em relação a características comportamentais, Mukherjee *et al*. (2013) identificaram que avaliadores maliciosos, em geral, não são membros ativos, de longa data, das plataformas de avaliação. Eles tendem a postar todas as suas opiniões em um único dia e as avaliações desviam da maioria das demais avaliações para um dado hotel. Nessa mesma linha, Wu *et al*. (2010) apresentam alguns critérios que podem ser indicativos de revisões suspeitas, como opiniões fortemente positivas imediatamente depois de uma opinião negativa ter sido postada, ou várias opiniões fortemente positivas emitidas uma após a outra, em um curto espaço de tempo. Por outro lado salientam que outras contribuições além de

texto, como fotos, vídeos ou itinerários de viagem, são um fator indicador de veracidade da avaliação.

3 METODOLOGIA

A fim de atingir os objetivos estabelecidos para esta pesquisa, a metodologia se concentrou em uma revisão sistemática da literatura, visando a analisar as formas apresentadas na literatura para formular um método de detecção de fraudes o mais completo possível. Com base na análise dos resultados dessa revisão, as características encontradas foram divididas entre as que ajudam a identificar o *spam* de propaganda e as que auxiliam na detecção do *spam* de difamação. Além dessa divisão, as características também foram separadas em comportamentais e linguísticas.

3.1 PROCEDIMENTOS ADOTADOS PARA REALIZAR A REVISÃO SISTEMÁTICA

Revisões sistemáticas da literatura são uma forma de identificar, avaliar e interpretar toda a pesquisa disponível relevante para uma determinada questão de pesquisa, área ou fenômeno de interesse (Kitchenham, 2004). Elas permitem incorporar um espectro maior de resultados relevantes, ao invés de limitar as conclusões à leitura de alguns artigos a que se tem acesso de forma não organizada (Mancini & Sampaio, 2007).

Para a revisão sistemática descrita nesse capítulo foram utilizados os procedimentos apresentados por Kitchenham (2004) e Mancini & Sampaio (2007), a partir das seguintes bases de dados: Google Scholar, Science Direct, Spring e portal de periódicos da Capes. Em todas as bases de dados *online*, foi estabelecido o seguinte critério, com o propósito de filtrar os artigos relevantes para a revisão sistemática da literatura:

- O artigo deve conter pelo menos uma das seguintes expressões: “*fake review*”, “*false review*”, “*review spam*” ou “*opinion spam*” em seu título, palavras-chave ou resumo.

Por conta da grande quantidade de itens encontrados, a filtragem dos artigos apresentados em cada base de dados ocorreu seguindo a ordem em que foram apresentados na tela e utilizando o critério de classificação por relevância. A busca era interrompida quando no mínimo cinco artigos irrelevantes para a pesquisa – que não atendiam ao critério de busca estabelecido – eram exibidos de forma sucessiva.

Após o término das buscas e da contagem de artigos selecionados em cada base, foi realizada a remoção de artigos repetidos, ou seja, contidos em mais de uma base ou que foram detectados a partir da busca por mais de uma expressão. A

Tabela 1, a seguir, mostra a quantidade de artigos encontrados a partir de cada base após a primeira etapa de filtragem.

Tabela 1. Artigos encontrados usando a filtragem por meio do critério de expressões.

Base de dados	Artigos encontrados
Capes	3
Google Scholar	15
Science Direct	6
Springer	7
Total	31

Fonte: autoria própria.

Os 31 artigos selecionados nessa etapa foram então submetidos a novos critérios de filtragem, cuja verificação foi realizada por meio da leitura do resumo e da introdução de cada artigo em sua íntegra:

- O artigo devia tratar da detecção de avaliações falsas em ambientes *online*.
- O artigo devia mostrar fatores que ajudassem a identificar as avaliações falsas.

Após filtragem dos artigos seguindo os critérios de filtragem mencionados acima, restaram quinze artigos. Foi então feita a leitura completa dos trabalhos que sobraram, certificando-se que eles realmente estavam relacionados ao que se buscava pesquisar, ou seja, discutiam formas de detectar falsas avaliações *online*. Nessa etapa foram eliminados mais seis artigos, restando nove artigos para serem incluídos no *corpus* da revisão sistemática.

Na Tabela 2, a seguir, são listados os critérios de exclusão dos 21 artigos cortados na análise do resumo, introdução e/ou texto integral.

Tabela 2. Critérios de exclusão dos artigos na análise do resumo e/ou texto integral.

Critérios de exclusão	**Artigos excluídos**
Artigos não disponíveis na íntegra na base	4
Artigos que não tinham como foco a detecção de avaliações falsas	11
Artigos que não identificavam características para auxiliar a detecção de avaliações falsas	6
Total	21

Fonte: autoria própria.

Quatro artigos não estavam disponíveis em sua íntegra nas bases consultadas. Antes da eliminação deles, tentou-se procurar em outra base de dados a que se tivesse acesso gratuito, mas não se obteve resultado e os artigos acabaram sendo descartados. Outros onze artigos discutiam avaliações em ambiente *online*, sem se preocupar com avaliações falsas e como identificá-las. Já os sete restantes, embora tivessem como tema principal a detecção de avaliações falsas *online*, não identificavam características que pudessem ajudar a identificar se uma avaliação é falsa.

3.2 PROCEDIMENTOS ADOTADOS PARA A ANÁLISE DOS RESULTADOS

Os nove artigos selecionados para compor o corpus da revisão sistemática da literatura estão relacionados no Apêndice A, com a indicação do ano de publicação e bases de dados onde foram encontrados. Os estudos são do período de 2007 a 2016, sendo que a maior parte deles foi publicada a partir de 2013. Isso mostra que o tema em questão vem sendo pesquisado com mais frequência apenas muito recentemente, provavelmente por ter sido percebido como um problema mais sério somente nos últimos anos.

A seguir, são apresentadas as características que indicam que uma avaliação pode ser falsa, encontradas em cada artigo selecionado. Elas são diferenciadas em características que identificam *spam* de propaganda (avaliações positivas falsas) e que identificam *spam* de difamação (avaliações negativas falsas). Além dessa diferenciação, elas também são separadas em características linguísticas e comportamentais. O quadro 1 mostra as características linguísticas que podem indicar *spam* de propaganda encontradas.

Quadro 1. Características linguísticas que podem indicar *spam* de propaganda.

Características linguísticas que podem indicar spam de propaganda	**Autor**
Avaliações positivas falsas utilizam superlativos.	Filieri (2016)
Avaliações fortemente positivas depois de uma negativa podem ser falsas.	Wu *et al.* (2010)
Avaliações superficiais tendem a ser falsas.	Filieri (2016)

Avaliações que só expressam sentimentos positivos podem ser falsas.	Lu *et al.* (2013), Yoo & Gretzel (2009) , Mukherjee *et al.* (2016), Crawford et al. (2015)
Avaliações positivas que tenham um estilo de *marketing*, com detalhes promocionais tendem a ser falsas.	Filieri (2016)

Fonte: autoria própria.

A partir da análise das características linguísticas para *spam* de propaganda dos artigos selecionados, notou-se que as avaliações positivas falsas usam superlativos em abundância, como “o melhor hotel” ou “incrível”. Elas também possuem um estilo de *marketing*, com detalhes promocionais que poderiam ser encontrados em um folheto ou no *site* do hotel. Avaliações superficiais, com detalhes irrelevantes que não fornecem detalhes sobre sua experiência do usuário com o hotel também são suspeitas. Na tentativa de se recuperar de uma avaliação negativa, a administração de um hotel pode reagir escrevendo várias avaliações positivas falsas, logo após ter recebido uma avaliação negativa.

Já com relação as características comportamentais que podem indicar *spam* de propaganda, o Quadro 2, a seguir, mostra o que foi mencionado na revisão sistemática da literatura.

Quadro 2. Características comportamentais que podem indicar *spam* de propaganda.

Características comportamentais que podem indicar spam de propaganda	**Autor**
Scores que diferem muito da classificação média.	Keates (2007), Mukherjee *et al.* (2013), Crawford et al. (2015)
Avaliador escreveu sobre apenas um hotel.	Keates (2007)

Fonte: autoria própria.

Já em relação a características comportamentais para *spam* de propaganda foram encontrados menos resultados, mas o fator mais importante é que o avaliador falso normalmente só realiza uma avaliação positiva, de um único hotel, e não tem costume de frequentar o *site* de compartilhamento de opiniões.

Com relação ao *spam* de difamação, o Quadro 3 mostra as características linguísticas que foram detectadas.

Quadro 3. Características linguísticas que podem indicar *spam* de difamação.

Características linguísticas que podem indicar spam de difamação	**Autor**
Avaliação menciona propriedade próxima a onde o usuário que dá a opinião se encontra como sendo superior.	Keates (2007)
Avaliações são curtas porque o avaliador não tem muito o que escrever.	Mukherjee *et al.* (2013), Filieri (2016), Crawford *et al.* (2015)
Avaliações só expressam sentimentos negativos.	Lu *et al.* (2013)
Avaliações semelhantes a outras avaliações anteriores.	Mukherjee *et al.* (2013) , Heydari *et al.* (2015), Crawford *et al.* (2015)
Inconsistência com as outras avaliações .	Mukherjee *et al.* (2016)

Fonte: autoria própria.

Com a análise das características linguísticas para *spam* de difamação dos artigos selecionados pela revisão sistemática, percebeu-se que as avaliações negativas falsas são mais curtas que o normal, porque o avaliador não tem muito o que escrever, já que provavelmente nunca visitou o hotel presencialmente (Mukherjee *et al*., 2013; Filieri, 2016). Por conta disso, ele tende a copiar textos de avaliações anteriores. A inconsistência entre uma avaliação e as demais é um forte fator de desconfiança. Como, por exemplo, se a maioria diz que o "hotel oferece acesso a wi-fi" e há uma avaliação em que o usuário diz "Internet é cobrada", essa avaliação deve ser considerada suspeita de ser falsa (Mukherjee *et al*., 2016). Por fim, o Quadro 4 mostra as características comportamentais identificadas que podem indicar spam de difamação.

Quadro 4. Características comportamentais que podem indicar *spam* de difamação.

Características comportamentais que podem indicar spam de difamação	**Autor**

Scores diferem muito da classificação média.	Keates (2007), Mukherjee *et al.* (2013), Crawford *et al.* (2015)
Avaliadores falsos não são membros ativos de longa data.	Mukherjee *et al.* (2013)
Avaliadores falsos postam várias avaliações negativas em um curto espaço de tempo.	Mukherjee *et al.* (2013), Lu *et al.* (2013), Crawford *et al.* (2015), Mukherjee *et al.* (2016)
Avaliações verdadeiras têm outras contribuições além de texto (fotos, vídeos e itinerários de viagem), algo que as falsas não possuem.	Wu *et al.* (2010), Filieri (2016)

Fonte: autoria própria.

Sobre as características comportamentais para *spam* de difamação, o fator mais relevante para distinguir uma avaliação verdadeira de uma falsa são as outras contribuições da avaliação além do texto. Imagens, vídeos, itinerários da viagem feitos pelo próprio avaliador indicam veracidade da informação e são algo que normalmente não se encontra nas avaliações negativas falsas (Wu *et al*., 2010; Filieri, 2016).

Com base nessas características levantadas por meio da revisão sistemática da literatura, será possível modelar um método para detectar o *spam* de propaganda e o *spam* de difamação.

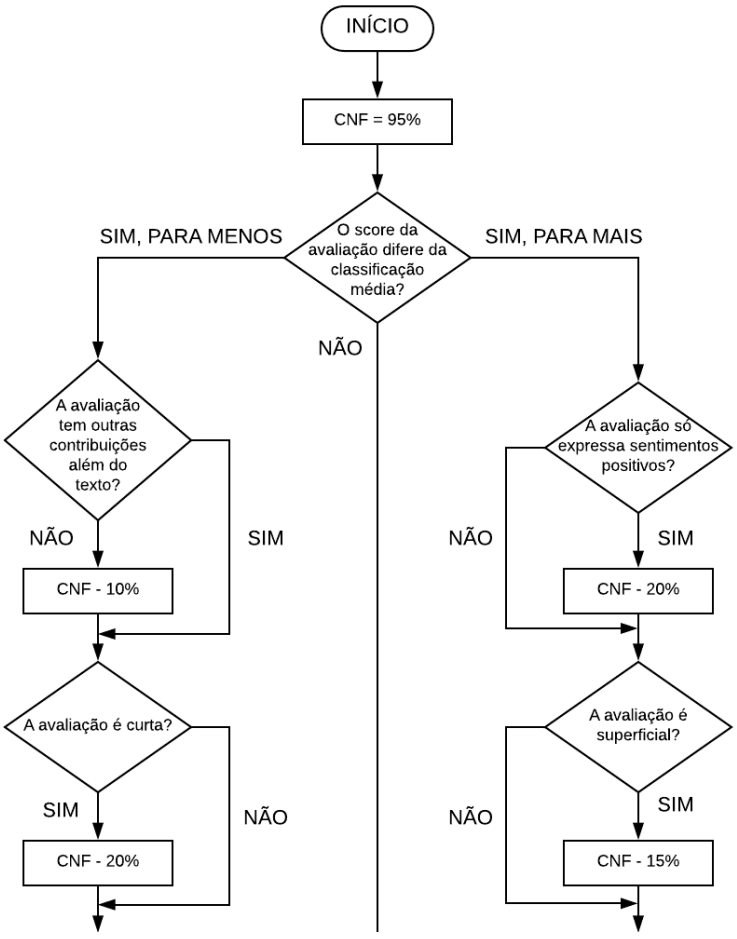
4 DESENVOLVIMENTO

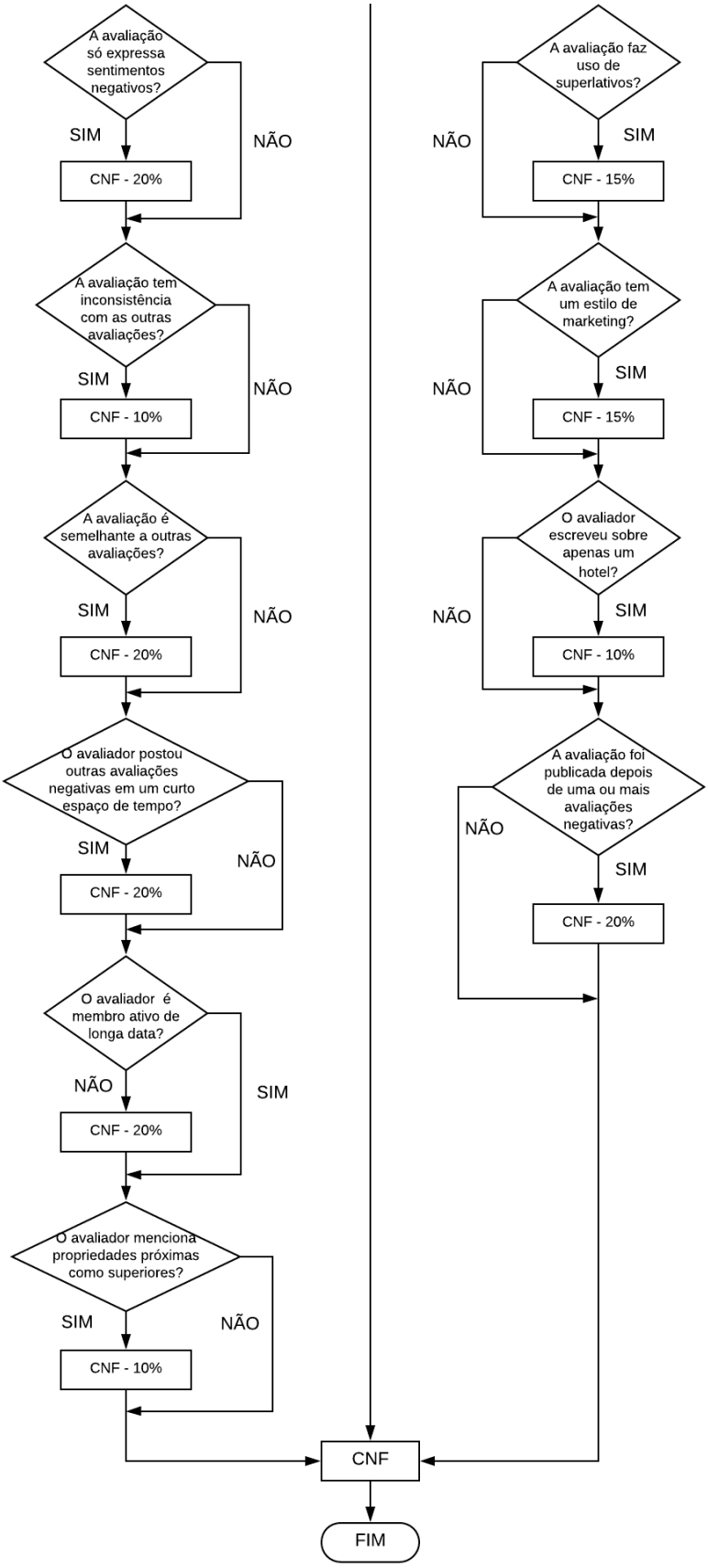
O checklist desenvolvido neste projeto tem a função de auxiliar na detecção de avaliações falsas de hotéis, *online*. Esse capítulo apresenta as etapas necessárias para o desenvolvimento, sendo dividido nas seguintes seções: 4.1 Modelagem, 4.2 Base de conhecimento e 4.3 Interface com o usuário.

4.1 MODELAGEM

Para auxiliar na criação da base de conhecimento de um checklist é recomendável passar por um processo de modelagem dos dados encontrados. Com base nos resultados encontrados na revisão sistemática da literatura, foi criado um fluxograma, utilizando-se o software Lucidchart. Essa ferramenta foi escolhida por conta da facilidade na criação de modelos. A Figura 1, a seguir, mostra o fluxograma desenvolvido para a implementação do checklist.

Figura 1. Fluxograma dos resultados da revisão sistemática da literatura.





Fonte: autoria própria.

Analisando os resultados obtidos nos estudos encontrados na revisão sistemática da literatura foi verificado que nenhuma característica linguística ou comportamental obteve 100% de exatidão para indicar se uma avaliação era falsa ou verdadeira. Sendo assim, foi definido que o checklist se iniciaria com uma CNF de 95% de chance de a avaliação ser verdadeira.

4.2 BASE DE CONHECIMENTO

Para a criação da base de conhecimento do Checklist foram declaradas as variáveis necessárias de acordo com as características encontradas na revisão sistemática da literatura, a ordem delas no sistema e seu grau de confiança. A variável inicial é "*O score* da avaliação difere da classificação média?". Keates (2007), Mukherjee et al. (2013) e Crawford (2015) mostraram em seus estudos que as classificações de spammers tendem a desviar-se da média a uma taxa muito mais alta do que os revisores legítimos. Isso acontece pelo fato que os spammers trabalham normalmente com extremos. Outro motivo para ela ser a primeira é que com a resposta dela podemos descobrir se estamos tentando detectar um spam de propaganda – quando a classificação desvia para cima - ou um spam de difamação – quando desvia para baixo. Caso a avaliação não desvie da média o fluxograma finaliza, com o resultado de 95% de chance de a avaliação ser verdadeira.

Os demais critérios não seguiram uma ordem especifica já que a posição de cada característica no fluxograma não interfere no resultado final. Essas variáveis possuem duas decisões de resposta: "sim" ou "não". Caso a resposta do critério aumentasse o nível de suspeita será subtraída uma porcentagem da CNF de acordo com o grau de certeza de cada característica. A tabela 3 a seguir mostra a subtração da CNF de cada varável do fluxograma.

Tabela 3. Subtração da CNF.

VARIÁVEL	VALORES	SUBTRAÇÃO DA CNF
A avaliação só expressa sentimentos positivos?	sim	CNF – 20 %
A avaliação é superficial?	sim	CNF – 15 %
A avaliação faz uso de superlativos?	sim	CNF – 15 %

A avaliação tem um estilo de marketing?	sim	CNF – 15 %
O avaliador escreveu sobre apenas um hotel?	sim	CNF – 10 %
A avaliação foi publicada depois de uma ou mais avaliações negativas?	sim	CNF – 20 %
A avaliação tem outras contribuições além do texto?	não	CNF – 10 %
A avaliação é curta?	sim	CNF – 20 %
A avaliação só expressa sentimentos negativos?	sim	CNF – 20 %
A avaliação tem inconsistência com as outras avaliações?	sim	CNF – 10 %
A avaliação é semelhante a outras avaliações?	sim	CNF – 20 %
O avaliador postou outras avaliações negativas em um curto espaço de tempo?	sim	CNF – 20 %
O avaliador é membro ativo de longa data?	não	CNF – 20 %
O avaliador menciona propriedades próximas como superiores?	sim	CNF – 10 %

Fonte: autoria própria.

A escolha da porcentagem a ser subtraída de cada variável foi decidida com base nos resultados dos estudos encontrados na revisão sistemática da literatura. As questões com base nas quais se subtraiu 20% da CNF, dependendo da resposta, são as que se referem as características apontadas em estudos quantitativos que indicam que tal resposta é um forte indicador para que a avaliação seja falsa. Por exemplo, no estudo de Mukherjee *et al.* (2013) foi verificado que a maioria das avaliações de *spammers* (80%) são limitadas a uma média de comprimento de 135 caracteres, o que é considerado curto se comparar com as avaliações verdadeiras que em 92% dos casos têm mais de 200 caracteres. Nesse mesmo estudo também foi verificado que 75% dos *spammers* escreveram mais de cinco resenhas em um só dia, enquanto 90% dos avaliadores legítimos nunca criam mais de uma decisão por dia.

Já as questões que levaram a subtrair 15% são de autores que realizaram estudos qualitativos, mas que utilizaram entrevistas com pessoas para descobrir a opinião delas a respeito das características. No estudo de Filieri (2016), os entrevistados mencionaram que avaliações indignas de confiança são superficiais, não fornecem informações factuais ou evidência de compra, fazem uso abundante de superlativos (linguagem emocional) e usam um estilo de escrita de "marketing".

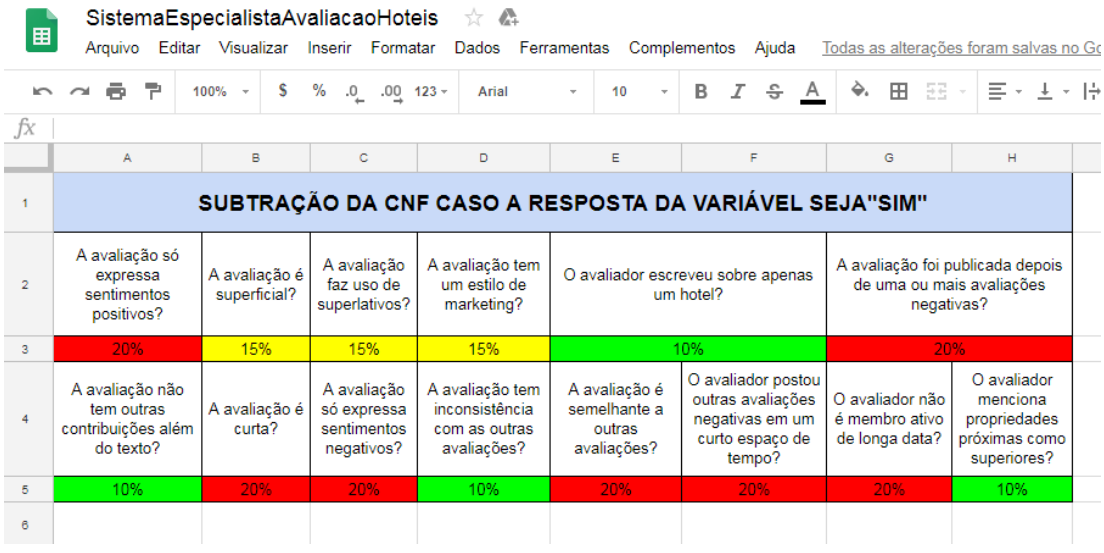
Por fim, para o restante das questões levantadas pelos autores da revisão sistemática da literatura subtraiu-se 10% da CNF no caso de respostas que aumentaram a suspeita de avaliação falsa.

4.3 INTERFACE COM O USUÁRIO

Para deixar o checklist mais intuitivo ao usuário e com uma interface simples, foi feita a escolha do Microsoft Office Excel para ele ser desenvolvido e utilizado.

O checklist é composto por duas planilhas. Na primeira delas estão listadas as porcentagens que serão subtraídas da CNF de cada variável cuja resposta aumenta o nível de suspeita. É possível editar as porcentagens caso se julgue necessário. A Figura 2, a seguir, mostra essa primeira planilha.

Figura 2. Primeira planilha do checklist.



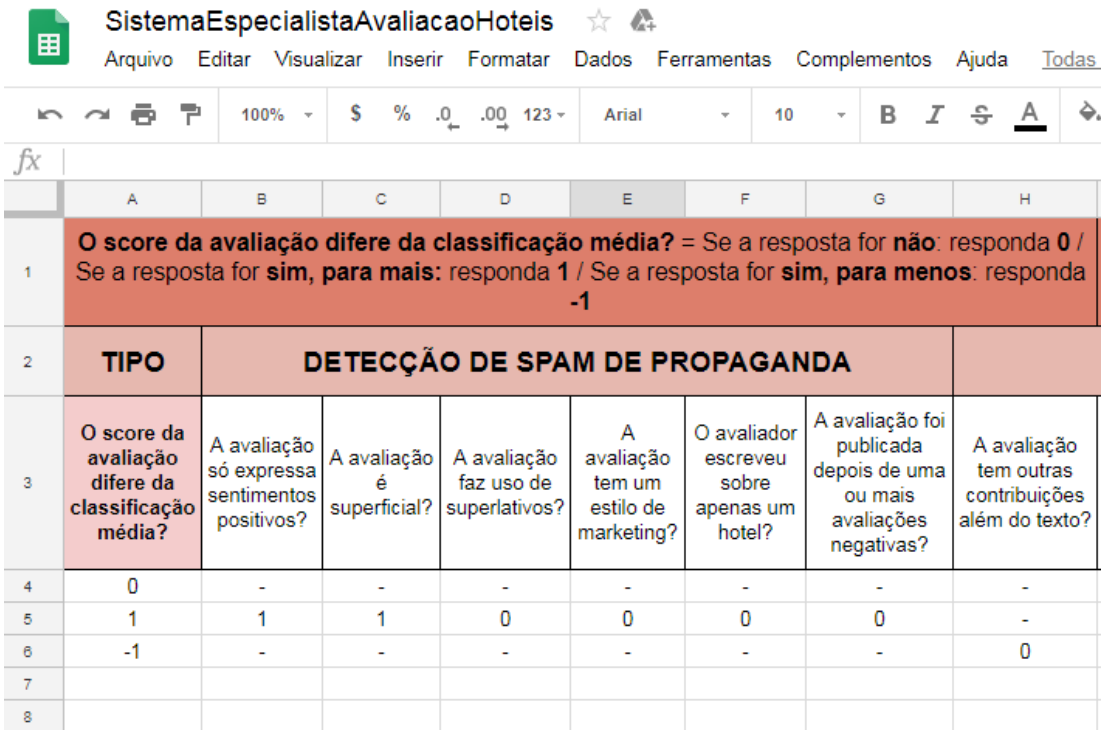
	A	B	C	D	E	F	G	H
1	SUBTRAÇÃO DA CNF CASO A RESPOSTA DA VARIÁVEL SEJA"SIM"							
2	A avaliação só expressa sentimentos positivos?	A avaliação é superficial?	A avaliação faz uso de superlativos?	A avaliação tem um estilo de marketing?	O avaliador escreveu sobre apenas um hotel?		A avaliação foi publicada depois de uma ou mais avaliações negativas?	
3	20%	15%	15%	15%	10%		20%	
4	A avaliação não tem outras contribuições além do texto?	A avaliação é curta?	A avaliação só expressa sentimentos negativos?	A avaliação tem inconsistência com as outras avaliações?	A avaliação é semelhante a outras avaliações?	O avaliador postou outras avaliações negativas em um curto espaço de tempo?	O avaliador não é membro ativo de longa data?	O avaliador menciona propriedades próximas como superiores?
5	10%	20%	20%	10%	20%	20%	20%	10%
6								

Fonte: autoria própria.

Já a segunda planilha é onde o usuário fornece as informações ao checklist, por meio de respostas às perguntas a respeito da avaliação a ser validada. Após receber as respostas das perguntas, o checklist retorna como resultado, o grau de

confiança de a avaliação ser verdadeira. As figuras 3 e 4, a seguir, mostram a segunda planilha.

Figura 3. Segunda planilha do checklist (parte 1).



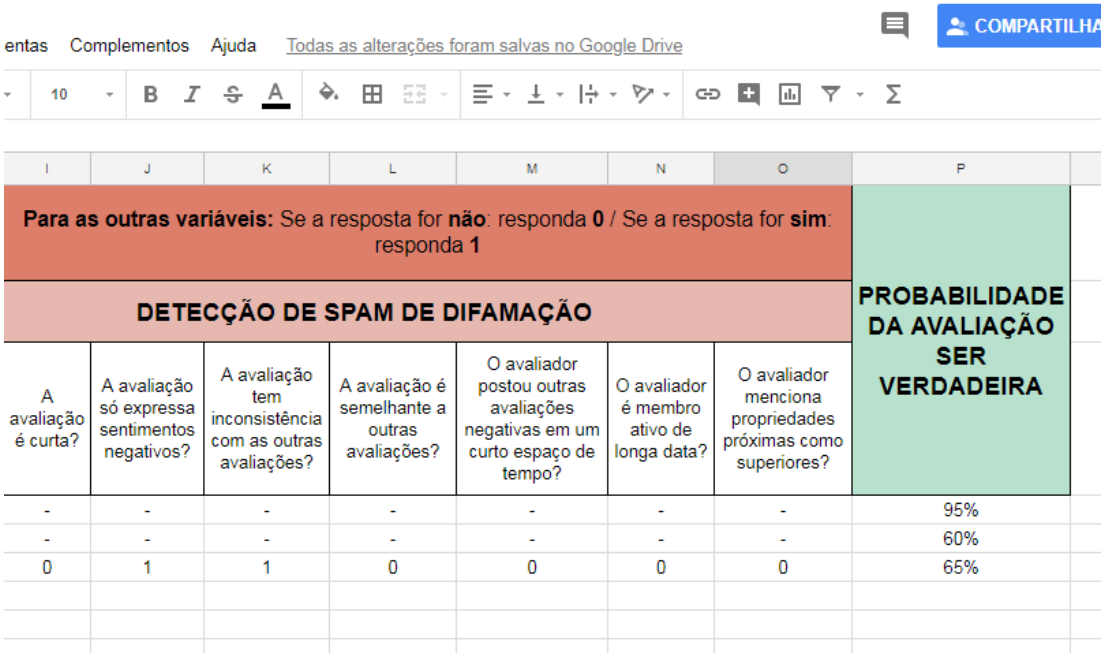
SistemaEspecialistaAvaliacaoHoteis

Arquivo Editar Visualizar Inserir Formatar Dados Ferramentas Complementos Ajuda Todas

	A	B	C	D	E	F	G	H
1	**O score da avaliação difere da classificação média?** = Se a resposta for **não**: responda **0** / Se a resposta for **sim, para mais:** responda **1** / Se a resposta for **sim, para menos**: responda **-1**							
2	**TIPO**	**DETECÇÃO DE SPAM DE PROPAGANDA**						
3	**O score da avaliação difere da classificação média?**	A avaliação só expressa sentimentos positivos?	A avaliação é superficial?	A avaliação faz uso de superlativos?	A avaliação tem um estilo de marketing?	O avaliador escreveu sobre apenas um hotel?	A avaliação foi publicada depois de uma ou mais avaliações negativas?	A avaliação tem outras contribuições além do texto?
4	0	-	-	-	-	-	-	-
5	1	1	1	0	0	0	0	-
6	-1	-	-	-	-	-	-	0
7								
8								

Fonte: autoria própria.

Figura 4. Segunda planilha do checklist (parte 2).



entas Complementos Ajuda Todas as alterações foram salvas no Google Drive COMPARTILHA

	I	J	K	L	M	N	O	P
	Para as outras variáveis: Se a resposta for **não**: responda **0** / Se a resposta for **sim**: responda **1**							
	DETECÇÃO DE SPAM DE DIFAMAÇÃO							**PROBABILIDADE DA AVALIAÇÃO SER VERDADEIRA**
	A avaliação é curta?	A avaliação só expressa sentimentos negativos?	A avaliação tem inconsistência com as outras avaliações?	A avaliação é semelhante a outras avaliações?	O avaliador postou outras avaliações negativas em um curto espaço de tempo?	O avaliador é membro ativo de longa data?	O avaliador menciona propriedades próximas como superiores?	
	-	-	-	-	-	-	-	95%
	-	-	-	-	-	-	-	60%
	0	1	1	0	0	0	0	65%

Fonte: autoria própria.

A primeira pergunta do checklist é “O *score* da avaliação difere da classificação média?”. Para essa pergunta o usuário possui três opções de resposta: 0 se a resposta for não, 1 se for sim, para mais e -1 se for sim, para menos. Para as demais perguntas o usuário tem duas opções de resposta: 1 se a resposta for sim ou 0 se for não.

5 VALIDAÇÃO DO SISTEMA

Para validar o checklist desenvolvido nesse trabalho foi necessário formular uma base de dados com avaliações de hotéis verdadeiras, falsas positivas (*spam* de propaganda) e falsas negativas (*spam* de difamação). Para não interferir nas plataformas de recomendação de hotéis, ao realizar nossos testes, foi pedido a voluntários que realizassem as avaliações de hotéis alimentando um formulário, como se estivessem realizando avaliações para um *site* como o TripAdvisor, ou outro semelhante. Esse capítulo apresenta as etapas necessárias para a validação, sendo dividido nas seguintes seções: 5.1 Formulário, 5.2 Limitações dos testes, 5.3 Análise dos resultados e 5.4 Aprimoramento do checklist.

5.1 FORMULÁRIO

Para a criação da base de dados para os testes, foi criado um formulário com questões semelhantes às que um usuário iria responder caso estivesse avaliando um hotel em uma plataforma de recomendação de hotéis. O formulário foi desenvolvido na plataforma Google Docs pela facilidade na divulgação e na armazenagem dos dados. Os voluntários, ao acessar o formulário, precisaram realizar três tipos de avaliações de hotéis:

- Uma avaliação franca e honesta de algum hotel no qual já tenha se hospedado.
- Uma avaliação falsa positiva de algum hotel que conheça ou no qual já tenha se hospedado. Nesse caso foi pedido que o voluntário agisse como se fosse o gerente desse hotel, escrevendo a avaliação com a intenção de promovê-lo positivamente.
- Uma avaliação falsa negativa de algum hotel, no qual nunca tenha se hospedado. Nesse caso foi pedido que o voluntário agisse como se fosse o gerente de um hotel concorrente ao hotel que iria avaliar, escrevendo com a intenção manchar sua reputação.

Caso o voluntário nunca tivesse se hospedado em um hotel ou não se achasse apto em fazer uma avaliação honesta sobre um, foi deixado como opcional a tarefa de realizar a avaliação verdadeira. As duas únicas avaliações obrigatórias eram a falsa positiva e a falsa negativa. O formulário em sua íntegra encontra-se no Apêndice B.

5.2 LIMITAÇÕES DOS TESTES

Pelo fato de a base de dados não ser originada por meio de uma plataforma de recomendações de hotéis, os testes para a validação do checklist tiveram limitações. Todas as características linguísticas puderam ser utilizadas durante os testes. Entretanto, somente a característica comportamental "*Scores* que diferem muito da classificação média." pode ser analisada nas avaliações realizadas pelos voluntários, por meio do formulário. Já as outras características que dizem respeito ao comportamento dos *spammers* dentro das plataformas não puderam ser contempladas nos testes, pois não era possível identifica-las no método de coleta de dados proposto. Essas características comportamentais, que não puderam ser alinhas, estão listadas a seguir:

- Avaliador escreveu sobre apenas um hotel.
- A avaliação foi publicada depois de uma ou mais avaliações negativas.
- Avaliações verdadeiras têm outras contribuições além de texto (fotos, vídeos e itinerários de viagem).
- Avaliadores falsos são membros ativos de longa data.
- Avaliadores falsos postam várias avaliações negativas em um curto espaço de tempo.

5.3 ANÁLISE DOS RESULTADOS

O formulário foi divulgado durante o período de 1 a 20 de abril de 2018, conseguindo a participação de 40 voluntários. Entre os participantes, todos colaboraram com uma avaliação falsa positiva (*spam* de propaganda) e outra falsa negativa (*spam* de difamação). No entanto, apenas 29 se consideraram aptos a escrever uma avaliação franca de um hotel em que se hospedou recentemente. Após uma filtragem das avaliações falsas, foram descartadas avaliações que não seguiam o que havia sido proposto e apenas consideradas para o teste 35 avaliações falsas positivas e 38 avaliações falsas negativas, além de todas as 29 avaliações verdadeiras.

Com as avaliações verdadeiras e falsas escritas pelos voluntários foi possível criar uma base de dados para avaliar a eficácia do checklist. Uma avaliação por vez foi analisada pelo checklist para detectar se tratava de uma avaliação verdadeira, falsa positiva ou falsa negativa. Foram consideradas para os testes duas porcentagens de CNF mínimas para presumir que a avaliação é verdadeira, 70% e 60%. O resultado dos testes é mostrado na Tabela 4, a seguir.

Tabela 4. Resultado da avaliação do checklist.

RESULTADO DA AVALIAÇÃO DO CHECKLIST		CNF mínima para considerar verdadeira	
		70%	**60%**
Avaliações verdadeiras (29)	Média de acertos	62.06%	82.75%
	Erros	11	5
	CNF média dos erros	49.54%	41%
Avaliação falsa: *spam* de propaganda (35)	Média de acertos	77.14%	37.14%
	Erros	8	22
	CNF média dos erros	78.12%	66.59%
Avaliação falsa: *spam* de dfamação (38)	Média de acertos	89.47%	50%
	Erros	4	19
	CNF média dos erros	75%	65%

Fonte: autoria própria.

O resultado mostra que o checklist conseguiu detectar a maior parte das avaliações verdadeiras nas duas porcentagens mínimas consideradas, mas houve uma dificuldade maior para detectar as falsas quando a CNF mínima foi reduzida para 60%. Isso aconteceu porque as avaliações que não conseguiram ser detectadas pelo checklist encontravam-se em sua maior parte próximo a CNF mínima, o que mostra a diferença significativa do número de erros entre as duas porcentagens mínimas consideradas.

Em relação as avaliações verdadeiras, a característica mais marcante observada foi que o score delas não destoa da classificação média. Foi possível identificar que maioria das avaliações verdadeiras na base de dados possuía um misto de sentimentos (negativos e positivos) e apresentavam detalhes que comprovava que a pessoa realmente se hospedou no hotel.

Apenas 5 avaliações verdadeiras continuaram tendo resultado *spam* de propaganda pelo checklist quando o grau de confiança mínimo foi reduzido para 60%. Todas elas apresentavam o *score* mais alto e somente havia sentimentos

positivos em seu texto. Isso mostra a dificuldade de detectar a veracidade de uma avaliação extrema positiva, mas que não tenha a intenção de promover artificialmente o hotel. Na base de dados não havia nenhuma avaliação verdadeira negativa sobre um hotel, então não foi possível validar o comportamento do checklist nesse caso. Contudo, é provável que também tenha essa dificuldade em detectar sua veracidade já que se supõe que alguém que se preste a entrar em um *site* de compartilhamento de opiniões para fazer uma avaliação sincera, mas negativa de um hotel tenha ficado muito pouco satisfeito com a experiência e, assim, não esteja voltado a salientar eventuais aspectos positivos, mesmo que houvesse. Algumas características comportamentais que não puderam ser contempladas no teste, como se o avaliador postou mais de uma avaliação na plataforma, talvez pudessem ajudar a garantir a lisura de um avaliador com esse tipo de avaliação.

O checklist teve maior dificuldade em detectar as avaliações falsas favoráveis ao hotel. Ele conseguiu reconhecer menos da metade delas como falsas (37,14%) quando considerada a CNF mínima de 60%. Um dos motivos desse resultado talvez seja a limitação de não utilizar a maioria das características comportamentais nos testes, já que, como dito anteriormente, ao analisar as avaliações verdadeiras, notou-se uma dificuldade em diferenciar as avaliações falsas de verdadeiras quando elas são extremamente positivas apenas a partir do seu conteúdo.

Das avaliações falsas favoráveis ao hotel que foram detectadas pelo checklist notou-se que elas são extremamente positivas, tendo o conteúdo com estilo de *marketing* e usando em abundância superlativos para expressar as qualidades do estabelecimento em grau máximo. Foi observado também, por meio dos testes, que elas eram normalmente superficiais, curtas, não possuindo detalhes que comprovassem que o avaliador realmente se hospedou no hotel.

Já em relação as avaliações falsas desfavoráveis, o checklist conseguiu detectar grande parte delas (89.47%), quando a CNF mínima considerada foi 70%, e apenas metade delas quando se reduzia para 60%. As características mais evidentes nas avaliações falsas que foram detectadas foi que elas só expressavam sentimentos negativos. Também havia inconsistência com outras avaliações, quando o texto desqualificava algum serviço do estabelecimento. Ao contrário das avaliações falsas favoráveis, notou-se que o *spam* de difamação consegue ser detectado mais facilmente apenas com as características linguísticas. Contudo, as características comportamentais que não foram contempladas no teste, como se o

avaliador postou outras avaliações negativas em um curto espaço de tempo ou se ele é um membro de longa data, ajudariam na detecção das avaliações falsas que não puderam ser identificadas a partir do seu texto. Essas avaliações que não conseguiram ser detectadas como falsas embora o fossem, eram normalmente longas e não mostravam inconsistências com outras avaliações.

5.4 APRIMORAMENTO DO CHECKLIST

Após a realização da análise dos resultados da validação do SE, foi verificada a quantidade de vezes que cada característica contemplada nos testes foi marcada com a alternativa que causava suspeita. Essa inspeção foi necessária para descobrir se a subtração da CNF poderia ser melhorada, aumentado o grau da porcentagem da CNF a ser subtraída para as características que obtiveram maior incidência e diminuindo, caso contrário, na intenção de aprimorar a precisão do checklist. O resultado dessa análise é mostrado na Tabela 5, a seguir.

Tabela 5. Análise das características e CNF.

ANÁLISE DAS CARACTERÍSTICAS E CNF		
Características	**Quantidade**	**Mudança na subtração da CNF em caso de suspeita**
A avaliação só expressa sentimentos positivos?	44	CNF - 20%
A avaliação é superficial?	26	CNF - 20%
A avaliação faz uso de superlativos?	15	CNF - 15%
A avaliação tem um estilo de *marketing*?	23	CNF - 15%
A avaliação é curta?	18	CNF - 15%
A avaliação só expressa sentimentos negativos?	35	CNF - 20%
A avaliação tem inconsistência com as outras avaliações?	31	CNF - 20%

A avaliação é semelhante a outras avaliações?	3	CNF - 10%
O avaliador menciona propriedades próximas como superiores?	1	CNF - 10%

Fonte: autoria própria.

Para verificar se a precisão do checklist seria melhorada com a mudança na subtração de CNF de cada característica, o checklist foi atualizado com os novos dados e foi novamente realizado o processo de validação. Nesse novo teste também foram consideradas as CNF mínimas de 70% e 60% para presumir que a avaliação é verdadeira. O resultado do teste com o checklist modificado é mostrado na Tabela 6, a seguir.

Tabela 6. Resultado da avaliação do checklist modificado.

RESULTADO DA AVALIAÇÃO DO CHECKLIST MODIFICADO		**CNF mínima para considerar verdadeira**	
		70%	**60%**
Avaliações verdadeiras (29)	Média de acertos	62.06%	82.75%
	Erros	11	5
	CNF média dos erros	49.54%	41%
Avaliação falsa: spam de propaganda (35)	Média de acertos	77.14%	65.71%
	Erros	8	12
	CNF média dos erros	78.12%	72.08%
Avaliação falsa: spam de difamação (38)	Média de acertos	89.47%	76.31%
	Erros	4	9
	CNF média dos erros	75%	67.22%

Fonte: autoria própria.

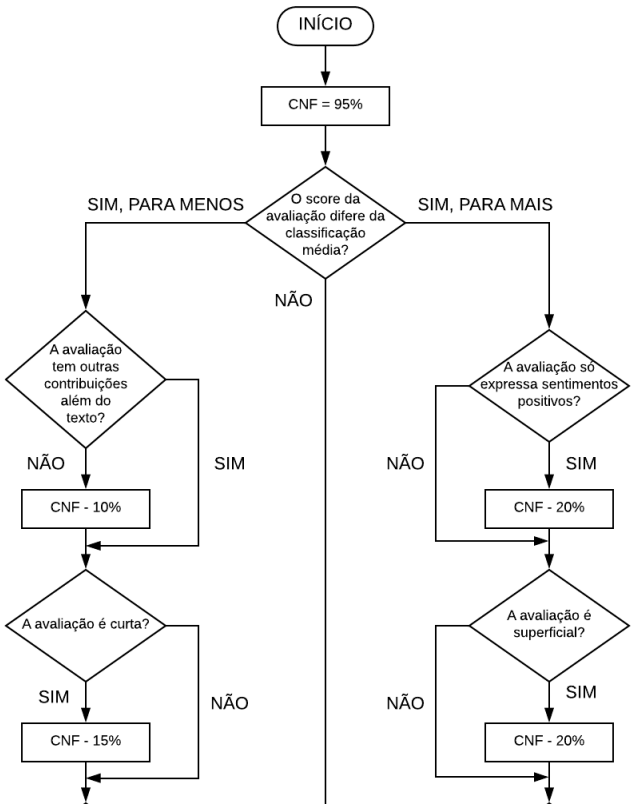
Não houve nenhuma mudança de resultado em relação as avaliações verdadeiras. Isso se deve ao fato de o grau de confiança das avaliações que foram

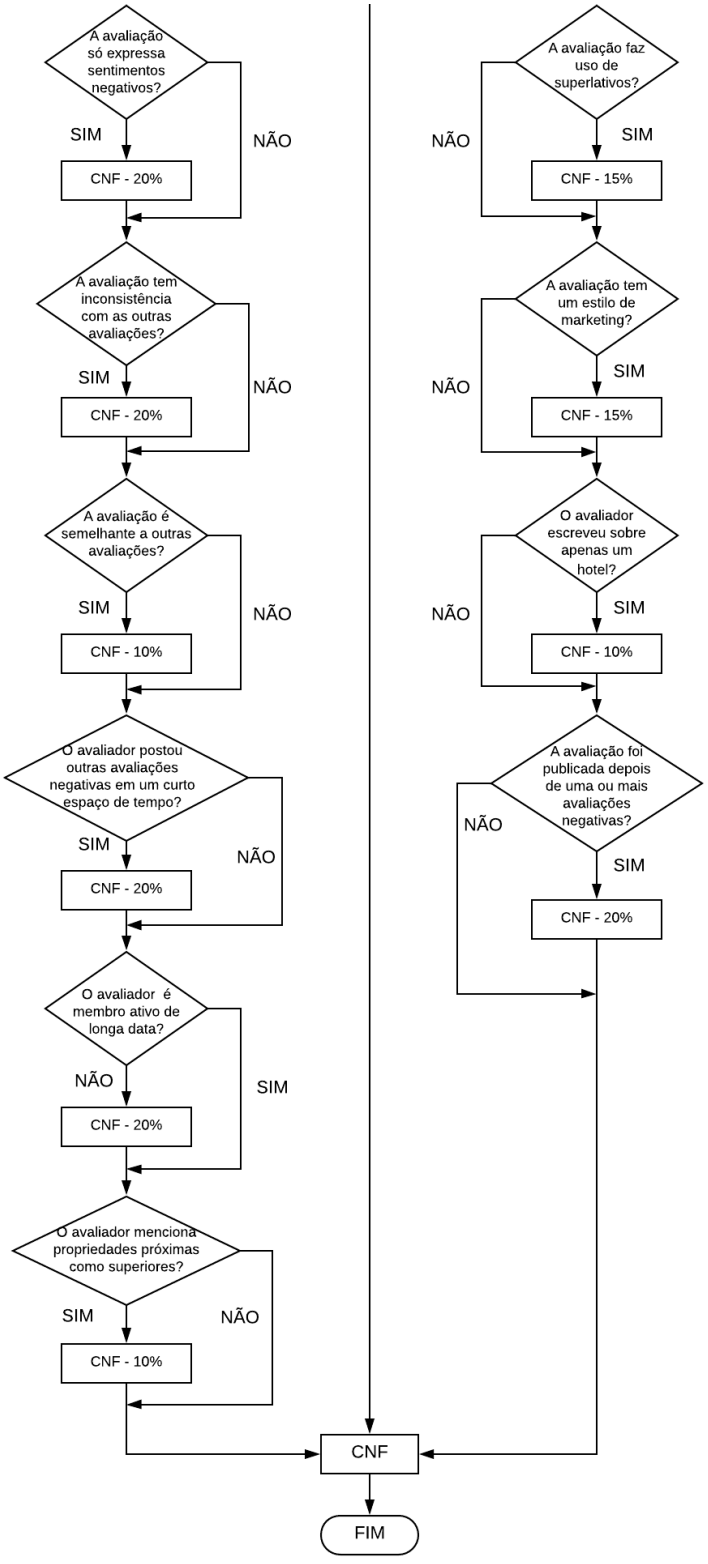
apontadas como *spam* de propaganda ser muito baixo, 41% quando a CNF mínima considerada foi 60%. Contudo, pela quantidade alta de acertos é possível declarar que o checklist atende às expectativas, já que é melhor desconsiderar uma quantidade mínima de avaliações verdadeiras do que acabar considerando avaliações falsas.

Em relação as avaliações falsas favoráveis e desfavoráveis não houve mudança nos resultados quando foi considerada a CNF mínima de 70%. Entretanto, houve uma alta diminuição da taxa de erros ao comparar com a validação anterior quando foi adotado 60% como valor mínimo de CNF. Isso se deu ao fato de aumentar a quantidade a ser subtraída das características mais fortes a serem analisadas, quando há a possibilidade de a avaliação ser falsa.

A partir do resultado obtido nessa validação pode-se garantir que o checklist foi aprimorado com as atualizações feitas. Assim, o fluxograma originado pela revisão sistemática da literatura foi remodelado para ser mais preciso em seus resultados. O fluxograma atualizado é apresentado na Figura 5, a seguir.

Figura 5. Fluxograma dos resultados da revisão sistemática da literatura atualizado.





Fonte: autoria própria.

6 CONSIDERAÇÕES FINAIS

Esse trabalho de conclusão de curso visou a investigar formas de detectar avaliações falsas em *sites* de compartilhamento de opinião sobre hotéis, desenvolvendo um checklist capaz de melhorar a capacidade de detecção desse tipo de avaliações. Críticas falsas para promover ou desacreditar um determinado produto/serviço, chamadas de *spam* de opinião, podem ser de dois tipos: *spam* de propaganda (avaliações falsas favoráveis) e *spam* de difamação (avaliações falsas desfavoráveis). Para detectar essas formas de *spam*, foi constatada a necessidade de encontrar características que distinguam entre avaliações verdadeiras e falsas. Essas características podem ser linguísticas (que têm foco no texto da crítica) ou comportamentais (concentram-se no comportamento do avaliador no *site*). Por meio de uma revisão sistemática da literatura, foram identificadas características já analisadas por outros pesquisadores. Elas foram separadas em 4 grupos:

- Características comportamentais que podem indicar *spam* de propaganda.
- Características linguísticas que podem indicar *spam* de propaganda.
- Características comportamentais que podem indicar *spam* de difamação.
- Características linguísticas que podem indicar *spam* de difamação.

Todas essas fases foram necessárias para ser possível realizar a modelagem de um checklist de detecção de avaliações falsas, no qual as regras de conhecimento do checklist foram desenvolvidas e depois validadas em testes que simulavam avaliações criadas em *sites* de compartilhamento de opiniões.

Os resultados da validação do checklist evidenciaram uma dificuldade já encontrada na revisão sistemática da literatura: a complexidade em distinguir se uma avaliação é verdadeira ou falsa. As características encontradas ajudam na detecção, mas não é possível garantir que a avaliação é verdadeira ou falsa a partir delas.

Também foi verificado, por meio dos testes que, o *spam* de propaganda é mais difícil de detectar que o *spam* de difamação, já que a porcentagem de detecção desse tipo de avaliação foi menor. O motivo de isso acontecer é o fato do *spammer* conhecer o estabelecimento que está avaliando podendo, assim, fornecer detalhes dos serviços prestados. Assim, esse tipo de avaliação falsa pode ser confundido com uma avaliação extrema positiva, mas que não tenha a intenção de promover artificialmente o hotel.

Já em relação ao *spam* de difamação, grande parte desse tipo de avaliação conseguiu ser detectada pelo checklist. Como o *spammer* não conhecia o estabelecimento que estava avaliando, as avaliações só apresentavam sentimentos negativos extremos e também inconsistência com outros textos publicados sobre o hotel.

Com a limitação dos testes em não poder avaliar a maioria das características comportamentais, notou-se que, para detectar o *spam* de propaganda, elas seriam importantes, reduzindo a chance de serem confundidas com avaliações extremas verdadeiras, enquanto o *spam* de difamação consegue, na maior parte das vezes, ser detectado apenas analisando o texto da avaliação.

Como trabalhos futuros, é possível pensar na utilização dessa pesquisa e do checklist para encontrar avaliações falsas em outros tipos de *sites* de compartilhamento de opinião, além de hotéis, como *e-commerce* e avaliações de estabelecimentos comerciais. Refazer a validação, utilizando as características comportamentais, seria importante para reavaliar o grau de precisão e a vulnerabilidade a falhas do sistema. Além disso, também pode-se desenvolver o checklist criado nesse projeto para um aplicativo, ajudando, assim, os hotéis a verificarem a veracidade das avaliações que estão sendo postadas sobre seu estabelecimento.

REFERÊNCIAS

BAMBAUER-SACHSE, Silke; MANGOLD, Sabrina. Do consumers still believe what is said in online product reviews? A persuasion knowledge approach. **Journal of Retailing and Consumer Services**, v. 20, n. 4, p. 373-381, 2013.

BANERJEE, Snehasish; CHUA, Alton Y. Applauses in hotel reviews: genuine or deceptive? In: **Science and Information Conference (SAI), 2014**. IEEE, p. 938-942, 2014.

BOND, Charles F.; DEPAULO, Bella M. Accuracy of deception judgments. **Personality and Social Psychology Review**, v. 10, n. 3, p. 214-234, 2006.

CHUA, A. Y.; BANERJEE, Snehasish. Reliability of reviews on the Internet: the case of Tripadvisor. In: **Proceedings of the World Congress on Engineering and Computer Science**. 2013.

CRAWFORD, M., KHOSHGOFTAAR, T. M., PRUSA, J. D., RICHTER, A. N., & Al NAJADA, H. Survey of review spam detection using machine learning techniques. **Journal of Big Data**, v.2, n. 1, p. 23, 2015.

FILIERI, Raffaele. What makes an online consumer review trustworthy? **Annals of Tourism Research,** v. 58, p. 46-64, 2016.

GRETZEL, Ulrike; YOO, Kyung Hyan. Use and impact of online travel reviews. **Information and Communication Technologies in Tourism**, v. 1, n. 1, p. 35-46, 2008.

HEYDARI, A.; TAVAKOLI, M.; SALIM, N.; & Heydari, Z. Detection of review spam: a survey. **Expert Systems with Applications**, v. 42, n. 7, p. 3634-3642, 2015.

JINDAL, Nitin; LIU, Bing. Opinion spam and analysis. In: **Proceedings of the 2008 International Conference on Web Search and Data Mining**. ACM, p. 219-230, 2008.

KEATES, Nancy. Deconstructing TripAdvisor. **Wall Street Journal**, v. 1, n. 4, 2007.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1-26, 2004.

LITVIN, Stephen W.; GOLDSMITH, Ronald E.; PAN, Bing. Electronic word-of-mouth in hospitality and tourism management. **Tourism Management**, v. 29, n. 3, p. 458-468, 2008.

LO, I. S.; MCKERCHER, B.; LO, A.; CHEUNG, C.; & LAW, R. Tourism and online photography. **Tourism Management**, v. 32, n. 4, p. 725-731, 2011.

LU, Y., ZHANG, L., XIAO, Y., & LI, Y. Simultaneously detecting fake reviews and review spammers using factor graph model. In **Proceedings of the 5th Annual ACM Web Science Conference**. ACM, p. 225-233, 2013.

LUCA, Michael; ZERVAS, Georgios. Fake it till you make it: reputation, competition, and Yelp review fraud. **Management Science**, v. 62, n. 12, p. 3412-3427, 2016.

MANCINI, M. C.; SAMPAIO, R. F. Estudos de revisão sistemática: um guia para síntese criteriosa da evidência científica. **Rev Bras Fisioter**, v. 11, n. 1, p. 83-89, 2007.

MAYZLIN, Dina; DOVER, Yaniv; CHEVALIER, Judith. Promotional reviews: an empirical investigation of online review manipulation. **The American Economic Review**, v. 104, n. 8, p. 2421-2455, 2014.

MUKHERJEE, A.; VENKATARAMAN, V.; LIU, B.; & GLANCE, N. Fake review detection: Classification and analysis of real and pseudo reviews. **Technical Report UIC-CS-2013–03, University of Illinois at Chicago, Tech. Rep.,** 2013.

MUKHERJEE, Subhabrata; DUTTA, Sourav; WEIKUM, Gerhard. Credible review detection with limited information using consistency features. In: **Joint European Conference on Machine Learning and Knowledge Discovery in Databases.** Springer International Publishing, p. 195-213, 2016.

O'CONNOR, Peter. User-generated content and travel: A case study on Tripadvisor. com. **Information and Communication Technologies in Tourism**, p. 47-58, 2008.

SCHINDLER, Robert M; BICKART, Barbara. Published word of mouth: referable, consumer-generated information on the Internet. **Online consumer psychology: understanding and influencing consumer behavior in the virtual world**, v. 32, 2005.

SPARKS, Beverley A.; BROWNING, Victoria. The impact of online reviews on hotel booking intentions and perception of trust. **Tourism Management**, v. 32, n. 6, p. 1310-1323, 2011.

WU, Guangyu; GREENE, Derek; CUNNINGHAM, Pádraig. Merging multiple criteria to identify suspicious reviews. In: **Proceedings of the Fourth ACM Conference on Recommender Systems**. ACM, p. 241-244, 2010.

YOO, Kyung-Hyan; GRETZEL, Ulrike. Comparison of deceptive and truthful travel reviews. **Information and Communication Technologies in Tourism**, v. 1, n. 1, p. 37-47, 2009.

ZHOU, Lina; SUNG, Yu-wei. Cues to deception in online Chinese groups. In: **Hawaii International Conference on System Sciences, Proceedings of the 41st annual**. IEEE, p. 146-146, 2008.

APÊNDICE A - RELAÇÃO DOS ARTIGOS SELECIONADOS PARA SEREM UTILIZADOS NA REVISÃO SISTEMÁTICA DA LITERATURA

Artigos selecionados na revisão sistemática da literatura	Base de dados	Ano
KEATES, Nancy. Deconstructing TripAdvisor. Wall Street Journal, v. 1, n. 4, 2007.	Google Scholar	2007
YOO, Kyung-Hyan; GRETZEL, Ulrike. Comparison of deceptive and truthful travel reviews. Information and Communication Technologies in Tourism, v. 1, n. 1, p. 37-47, 2009.	Google Scholar, Springer, Science Direct, Capes	2009
WU, Guangyu; GREENE, Derek; CUNNINGHAM, Pádraig. Merging multiple criteria to identify suspicious reviews. In: Proceedings of the Fourth ACM Conference on Recommender Systems. ACM, p. 241-244, 2010.	Google Scholar, Springer	2010
LU, Y., ZHANG, L., XIAO, Y., & LI, Y. Simultaneously detecting fake reviews and review spammers using factor graph model. In Proceedings of the 5th Annual ACM Web Science Conference. ACM, p. 225-233, 2013.	Google Scholar, Springer	2013
MUKHERJEE, A.; VENKATARAMAN, V.; LIU, B.; & GLANCE, N. Fake review detection: Classification and analysis of real and pseudo reviews. Technical Report UIC-CS-2013–03, University of Illinois at Chicago, Tech. Rep., 2013.	Google Scholar, Springer, Science Direct, Capes	2013
CRAWFORD, M., KHOSHGOFTAAR, T. M., PRUSA, J. D., RICHTER, A. N., & Al NAJADA, H. Survey of review spam detection using machine learning techniques. Journal of Big Data, v.2, n. 1, p. 23, 2015.	Springer, Google Scholar	2015
HEYDARI, A.; TAVAKOLI, M.; SALIM, N.; & Heydari, Z. Detection of review spam: a survey. Expert Systems with Applications, v. 42, n. 7, p. 3634-3642, 2015.	Capes	2015
FILIERI, Raffaele. What makes an online consumer review trustworthy? Annals of Tourism Research, v. 58, p. 46-64, 2016.	Science Direct, Google Scholar	2016
MUKHERJEE, Subhabrata; DUTTA, Sourav; WEIKUM, Gerhard. Credible review detection with limited information using consistency features. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer International Publishing, p. 195-213, 2016.	Springer, Google Scholar, Science Direct, Capes	2016

APÊNDICE B – FORMULÁRIO

Detecção de Avaliações Falsas de Hotéis Online

É cada vez mais comum que as pessoas leiam as experiências de outros hóspedes de hotéis na Internet e também escrevam e compartilhem suas próprias opiniões, para se ajudarem mutuamente a tomar melhores decisões sobre onde se hospedar. Contudo, nem todos os comentários nas plataformas de recomendação de hotéis são, necessariamente verdadeiros, porque há interesses em jogo e pessoas inescrupulosas podem tentar modificar o panorama a seu favor.

Assim, estamos desenvolvendo um trabalho cujo principal objetivo é desenvolver um sistema especialista capaz de melhorar a detecção de avaliações falsas em sites de compartilhamento de opinião sobre hotéis, tentando garantir a autenticidade das opiniões neles compartilhadas.

Para avaliar o grau de precisão do sistema especialista criado é necessário dispor de uma base de dados com avaliações verdadeiras e falsas e é nesse sentido que gostaríamos de poder contar com a sua colaboração.

Como não queremos interferir no funcionamento das próprias plataformas de recomendação ao realizarmos nossos testes, porque estaríamos causando o mesmo tipo de problema que a ação das pessoas inescrupulosas que, justamente, queremos evitar, estamos solicitando aos nossos participantes voluntários que realizem três avaliações de hotéis aqui mesmo nesse formulário, mas imaginando que as estivessem realizando para o TripAdvisor ou outro site semelhante.

Muito obrigado pela participação!

*Obrigatório

1) Escreva uma avaliação franca e honesta de algum hotel no qual já se hospedou.

Nome do Hotel:

Sua resposta

Cidade:

Sua resposta

Sua classificação para o hotel:

	1	2	3	4	5	
menor nota	○	○	○	○	○	maior nota

Escreva sua avaliação de forma textual, destacando os motivos por que gostou ou não gostou do hotel.

Sua resposta

2) Escreva uma avaliação falsa positiva de algum hotel que conheça ou no qual já se hospedou. Uma avaliação falsa positiva é também chamada de spam de propaganda, que são opiniões positivas não merecedoras de crédito para promover um produto/serviço. No caso de avaliações de hotéis, elas são feitas pelo próprio hotel ou com o apoio/incentivo dele para promovê-lo positivamente.

Observações:

- Proceda como se fosse o gerente desse hotel, escrevendo essa avaliação com a intenção de promovê-lo positivamente.

Nome do Hotel: *

Sua resposta

Cidade: *

Sua resposta

Sua classificação para o hotel: *

	1	2	3	4	5	
menor nota	○	○	○	○	○	maior nota

Escreva sua avaliação de forma textual, destacando os motivos por que teria gostado de se hospedar no hotel. *

Sua resposta

Se você usou o TripAdvisor ou outro site para buscar informações sobre o hotel para o qual fez a avaliação falsa positiva, coloque o link aqui:

Sua resposta

3) Escreva uma avaliação falsa negativa de algum hotel, no qual nunca tenha se hospedado. Uma avaliação falsa negativa é também chamada de spam de difamação, que são opiniões negativas injustas, maliciosas ou falsas para prejudicar um produto/serviço. No caso de avaliações de hotéis, elas são feitas por concorrentes com a intenção de prejudicar a reputação de outra empresa no mercado.

Observações:

- Proceda como se fosse o gerente de um hotel concorrente a esse que você está avaliando, escrevendo com a intenção manchar sua reputação.

Nome do Hotel: *

Sua resposta

Cidade: *

Sua resposta

Sua classificação para o hotel: *

	1	2	3	4	5	
menor nota	○	○	○	○	○	maior nota

Escreva sua avaliação de forma textual, destacando os motivos por que não teria gostado de se hospedar no hotel. *

Sua resposta

Se você usou o TripAdvisor ou outro site para buscar informações sobre o hotel para o qual fez a avaliação falsa negativa, coloque o link aqui:

Sua resposta