UNIVERSIDADE NOVE DE JULHO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA
E GESTÃO DO CONHECIMENTO

HUOSTON RODRIGUES BATISTA

FRAMEWORK PARA MINERAÇÃO DE OPINIÕES EM MÍDIAS SOCIAIS PARA DESCOBERTA DE CONHECIMENTO DO CLIENTE

São Paulo
2017

HUOSTON RODRIGUES BATISTA

FRAMEWORK PARA MINERAÇÃO DE OPINIÕES EM MÍDIAS SOCIAIS PARA DESCOBERTA DE CONHECIMENTO DO CLIENTE

Dissertação apresentada ao Curso de Mestrado em Informática e Gestão do Conhecimento da Universidade Nove de Julho como requisito parcial à obtenção do título de Mestre em Informática e Gestão do Conhecimento.

Prof. Dr. Marcos Antonio Gaspar (Orientador)

Prof. Dr. Renato José Sassi (Coorientador)

São Paulo

2017

FICHA CATALOGRÁFICA

Batista, Huoston Rodrigues.
Framework para mineração de opiniões em mídias sociais para descoberta de conhecimento do cliente. / Huoston Rodrigues Batista. 2017.
76 f.

Dissertação (Mestrado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2017.
Orientador (a): Prof. Dr. Marcos Antonio Gaspar.

1. Mineração de dados. 2. Mineração de textos. 3. Mineração de opiniões. 4. Conhecimento do cliente. 5. Redes sociais.
I. Gaspar, Marcos Antonio. II. Titulo.

CDU 004

UNINOVE
Universidade Nove de Julho

PARECER – EXAME DE DEFESA

Parecer da Comissão Examinadora designada para o exame de defesa do Programa de Pós-Graduação em Informática e Gestão do Conhecimento a qual se submeteu o aluno **Huoston Rodrigues Batista**

Tendo examinado o trabalho apresentado pelo aluno **Huoston Rodrigues Batista**, aluno regularmente matriculado no Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE, para obtenção do título de "Mestre em Informática e Gestão do Conhecimento", com dissertação intitulada ": Framework para mineração de opiniões em mídias sociais para descoberta de conhecimento do cliente", e após ter ouvido a exposição do candidato, a Comissão o considerou:

(X) **Aprovado** ( ) **Aprovado condicionalmente**

( ) **Reprovado com direito a novo exame** ( ) **Reprovado**

EXAMINADORES

Prof. Dr. Marcos Antonio Gaspar

Prof. Dr. Leandro Augusto da Silva

Prof. Dr. Renato Jose Sassi

São Paulo, 15 de dezembro de 2017.

Dedico este trabalho à minha família, em especial à Denise, que sempre me apoiou e acreditou em mim, sobretudo nos momentos de maior dificuldade.

AGRADECIMENTOS

Tenho muitos agradecimentos a fazer, mas sem dúvida, devo começar por quem mais me apoiou desde o começo desta jornada, enfrentando comigo todas as dificuldades pelas quais passei: Silvia Denise. Mais que uma companheira, foi a responsável por meu ingresso na carreira científica e segue dividindo comigo não apenas uma vida, mas uma visão de futuro.

Devo agradecimentos aos meus orientadores, a começar pelo Prof. Marcírio Chaves, meu primeiro orientador, que me ensinou sobre o rigor da pesquisa científica, e a quem muito admiro. Agradeço também ao Prof. Marcos Vinicius, meu segundo orientador, que me conduziu por boa parte do caminho, sempre com muita paciência, e de quem tornei-me amigo. Agradeço especialmente ao Prof. Marcos Gaspar, meu último orientador, com quem aprendi muito, não somente sobre pesquisa científica, e em quem me inspiro constantemente para tentar ser não somente um bom professor, mas também um ser humano melhor. Sua humildade e sabedoria serão sempre um ideal de vida para mim. Por último, mas não menos importante, agradeço ao Prof. Renato Sassi, meu coorientador por sua sabedoria e paciência comigo, e por seus conselhos e observações sempre precisos e extremamente valiosos, sem as quais esta pesquisa não teria sido possível. Agradeço imensamente à oportunidade de ter sido acolhido pelos professores Marcírio, Gaspar e Ivanir. Prometi não os decepcionar e sigo me esforçando para cumprir esta promessa.

Agradeço aos amigos que fiz durante minha caminhada no mestrado e que me ajudaram em momentos tão difíceis para alguém com minhas inúmeras limitações. Cito em especial Fabio Kazuo Ohashi, Fabio Falchi Magalhães e José Carmino Gomes Junior, por suas contribuições generosas. Aos amigos e colegas quem não citei, mas moram em meu coração, meu muito obrigado.

Por fim, agradeço ao carinho de minha família, em especial minha mãe, que mesmo longe, participou ativamente de minha rotina, e meu pai, cujos conselhos sempre me ajudaram a encarar desafios de forma otimista, mesmo quando não era fácil fazê-lo.

À Universidade Nove de Julho, meu agradecimento pelo investimento feito em minha formação, sem o apoio da qual eu jamais teria conseguido chegar onde cheguei. À secretaria de pós-graduação da UNINOVE pelo apoio recebido durante todo o curso.

“Inteligência é a capacidade de se adaptar à mudança.”

(Stephen Hawking).

RESUMO

Com a disseminação da Internet e popularização de tecnologias móveis, as relações entre clientes e empresas sofreram transformações. Comentários em relação a empresas, produtos ou serviços, antes restritos aos círculos de amizade, agora são compartilhados de forma constante e prolífica em redes sociais e sites especializados em receber opiniões de clientes em relação às suas experiências. Este fenômeno proporciona oportunidades para descoberta de conhecimento a partir destas opiniões, mas também desafios, considerando-se que, dada sua natureza e forma, as opiniões dos clientes consistem em dados não estruturados, que por sua vez demandam tratamentos específicos. Esta pesquisa tem por objetivo apresentar um *framework* para mineração de opiniões visando a descoberta de conhecimento do cliente em relação às suas experiências em empresas (restaurantes), com base em dados não estruturados extraídos de redes sociais, aplicável à realidade de pequenas e médias empresas. A rede social abordada nesta pesquisa foi o TripAdvisor, de onde foram extraídos dados de quatro empresas (restaurantes) por meio da técnica de *web scraping*. Os dados da primeira empresa foram usados para desenvolver e refinar o *framework*, que por sua vez, foi aplicado aos dados das demais. Estes dados foram submetidos a técnicas de mineração de textos como Análise de Sentimentos e Modelagem de Tópicos por meio da abordagem *tidy data* tais quais, tokenização, normalização, remoção de *stop words*, remoção de caracteres especiais e números, criação de bi-gramas, cálculo de pesos dos termos, comparações e contagens. Como principais resultados, destaca-se a geração de sumarizações e visualizações gráficas que contribuíram para evidenciar conhecimento acerca das relações entre diversas expressões e termos que não eram óbvias. Estas, por sua vez, foram descobertas a partir das análises efetuadas, que permitiram encontrar relações latentes entre termos citados por diferentes clientes. A Análise de Sentimentos aliada à Modelagem de tópicos revelou que os aspectos mais abordados pelos clientes se referem à comida, ao lugar e o atendimento, variando em intensidade e polaridade. A contribuição prática deste trabalho reside na aplicação da Mineração de Textos para revelar padrões e possibilitar a descoberta de conhecimento a partir das opiniões de clientes extraídas de redes sociais. O *framework* empregado provou-se útil como ferramenta para compreender melhor o cliente, suas expectativas e até mesmo suas frustrações, gerando assim conhecimento acerca dos clientes para benefício da empresa.

Palavras-chave: Mineração de Dados. Mineração de Textos. Mineração de Opiniões. Conhecimento do Cliente. Redes Sociais.

ABSTRACT

With the spread of the Internet and popularization of mobile technologies, relations between customers and businesses have been transformed. Comments about the company, products or services, previously restricted to circles of friendship, now are shared consistently and prolifically on social networks and websites specializing in receiving opinions from customers regarding their experiences. This phenomenon provides opportunities for knowledge discovery from these opinions, but also challenges, considering that, given their nature and form, customer reviews consist of unstructured data, which in turn require specific treatments. This research aims to present a opinion mining framework for customer knowledge discovery in relation to their experiences in restaurants, based on unstructured data extracted from social networks, applicable to the reality of Small and Medium Enterprises. The social network chosen for the development of this research was TripAdvisor, from which data were extracted from four restaurants through the technique of web scraping. The data of the first company were used to develop and refine the framework, which in turn, was applied to the data of the other companies. The data were processed through a series of text mining techniques, including Sentiment Analysis and Topic Modeling using the tidy data approach, such as tokenization, normalization, removal of stop words, removal of special characters and numbers, creation of bi-grams, calculation of relevance of terms, comparisons and counts. As main results, we highlight the generation of summaries and graphic visualizations that contributed to evidence knowledge about the relations between several expressions and terms that were not obvious. These, in turn, were discovered from the analysis made, which allowed finding latent relationships between terms cited by different customers. The Sentiment Analysis allied to the Topic Modeling revealed that the aspects most addressed by the clients refer to the food, the place, and the service, varying in intensity and polarity. The practical contribution of this work lies in the application of Text Mining to reveal patterns and enable the discovery of knowledge from the opinions of customers extracted from social networks. The framework proposed and applied in this research proved useful as a tool to better understand the client, his expectations, and even his frustrations, thus generating knowledge about the clients for the benefit of the company.

Keywords: Data Mining. Text Mining. Opinion Mining. Customer Knowledge. Social Networks.

LISTA DE ABREVIAÇÕES

CK	*Customer Knowledge*
CKM	*Customer Knowledge Management*
CRM	*Customer Relationship Management*
CSV	*Comma-separated values*
DTM	*Document Term Matrix*
GC	Gestão do Conhecimento
HTML	*HyperText Markup Language*
IDF	*Inverse document frequency*
LDA	*Latent Dirichlet Allocation*
LSA	*Latent Semantic Analysis*
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
pLSA	*Probabilistic Latent Semantic Analysis*
PME	Pequenas e Médias Empresas
SEBRAE	Serviço Brasileiro de Apoio às Micro e Pequenas Empresas
SVM	*Support Vector Machine*
TF	*Term frequency*
TF-IDF	*Term frequency–inverse document frequency*
TM	*Text Mining*

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

No atual ambiente competitivo no qual as empresas estão inseridas, a promoção da interação entre os recursos e conhecimentos da empresa é vital para o seu sucesso. Não há meio de comunicação mais rápido ou interativo que a internet (KAPLAN; HAENLEIN, 2010; KIETZMANN *et al.*, 2011). Os clientes podem entrar em contato com as empresas e outros clientes por meio de redes sociais de forma mais interativa do que nunca, permitindo assim às empresas aprofundarem suas relações com os clientes. Dessa forma, conteúdos e experiências gerados pelo cliente têm dominado as implementações da comunicação corporativa com o cliente em redes sociais (KIETZMANN *et al.*, 2011).

Baixo custo, personalização e facilidade de criação de conteúdos focada em mensagens através de mídias sociais apresentam-se como vantagens relevantes sobre os canais de comunicação tradicionais (HOFFMAN; FODOR, 2010). Tal contexto torna o uso das mídias sociais não apenas relevante, mas um fator estratégico para as empresas, independentemente de seu tamanho ou segmento de atuação. Entretanto, utilizar as mídias sociais não é uma tarefa fácil e pode exigir novas formas de pensar por parte das empresas (KAPLAN; HAENLEIN, 2010).

Essas características do atual ambiente de negócios tornam as mídias sociais um meio particularmente proveitoso para Pequenas e Médias Empresas (PMEs), segmento conhecido tanto por suas limitações de recursos, dado que este tipo de negócio muitas vezes não possui gerenciamento de suas áreas de negócios e não investem em ferramentas administrativas ou de relacionamento com seus clientes, quanto por seu potencial (LIN, 2014).

No Brasil, as PMEs respondem por 27% do PIB nacional, 52% dos empregos com carteira assinada e 40% dos salários pagos, segundo dados do SEBRAE - Serviço Brasileiro de Apoio às Micro e Pequenas Empresas (SEBRAE, 2014). Tal participação é relevante e justifica a preocupação com o desempenho e potencial evolução das empresas inseridas neste segmento em especial, bem como de seu principal ativo: o cliente.

Considerando-se a importância do cliente para as PMEs, o conhecimento dele proveniente é de extrema relevância, tendo em vista que numa economia cada vez mais baseada no conhecimento, os clientes estão mais capacitados, mais conectados e mais informados, como consequência natural dos avanços tecnológicos e do acesso generalizado à internet e outros meios de comunicação (PRAHALAD; RAMASWAMY, 2004).

Em função do contexto apresentado, o conhecimento do cliente tem despertado o interesse de pesquisadores enquanto disciplina e, além disso, como fonte estratégica fundamental para o sucesso de qualquer empresa (ROWLEY, 2002; CAMPBELL, 2003; ROLLINS; HALINEN, 2005). Isto se dá seja com o intuito de proporcionar inovação, facilitar a abertura de novos mercados e oportunidades de negócios ou suportar o gerenciamento de longo prazo com os clientes (DARROCH; MCNAUGHTON, 2003).

Entretanto, o conhecimento, como importante recurso para as empresas contemporâneas, advém muitas vezes de dados e informações que precisam ser processados a fim de assumir novas configurações que transmitam significado, contribuindo assim, para a tomada de decisões na empresa.

Com a evolução tecnológica marcada pela conectividade e popularização de dispositivos móveis e, sobretudo, pela penetração cada vez maior da Internet em todas as esferas da sociedade moderna, constata-se um aumento cada vez mais expressivo do volume de dados. Tal elevação no volume mencionado provém, predominantemente de dados não estruturados, tais como imagens, vídeos e, principalmente, textos publicados em redes sociais (SILVA; PERES; BOSCARIOLI, 2017).

Torna-se necessário, portanto, desenvolver técnicas e métodos que possibilitem extrair conhecimento destes dados, tarefa esta que, dada sua complexidade, motivou o surgimento de diferentes métodos e técnicas para realização de mineração de dados provenientes da complexa teia de repositórios das redes sociais.

Esta pesquisa apresenta um *framework* para mineração de opiniões que possa ser aplicado na descoberta de conhecimento do cliente em relação às suas experiências, com base em dados não estruturados extraídos de redes sociais, e que seja aplicável à realidade de pequenas e médias empresas.

O intuito do *framework* proposto é apresentar uma forma eficiente, rápida e econômica para extrair, pré-processar e minerar dados, apresentando os resultados da descoberta de conhecimento a partir da utilização de técnicas como a Análise de Sentimentos e Modelagem de Tópicos por meio de diferentes formas de visualizações. Dados provenientes de redes sociais de quatro empresas (restaurantes) foram extraídos e usados nesta pesquisa. Os dados de uma delas (a de menor quantidade de opiniões) foram usados para a construção e refinamento do *framework*. Em seguida, o *framework* elaborado foi aplicado aos dados das demais empresas como forma de demonstrar sua viabilidade como método e ferramental empregados, além de sua eficiência em termos práticos para as PMEs.

A principal contribuição que se pretende com esta pesquisa é a aplicação de uma forma simples e eficiente de extrair, tratar e apresentar o conhecimento proveniente do cliente em relação às suas experiências, tendo como base dados extraídos de redes sociais de empresas, considerando sua aplicação à realidade de PMEs.

1.1 Motivação

No Brasil, assim como em muitos outros países, a maioria das empresas é de pequeno e médio porte, sendo que o papel destas na economia do país é extremamente impactante para o crescimento econômico, emprego e geração de riquezas (LIN, 2014).

O universo de empresas que se enquadram no perfil de Pequenas e Médias Empresas (PME) no Brasil é de aproximadamente 8,9 milhões (SEBRAE, 2014). Há várias classificações para definir Pequenas e Médias Empresas. O principal critério que países da OCDE - Organização para a Cooperação e Desenvolvimento Econômico (2004) aplicam é o número de empregados. No Quadro 1 apresenta-se o indicador mais utilizado no Brasil, por sua vez baseado na classificação do SEBRAE (2014), e utilizado como critério neste estudo.

Quadro 1 – Classificação de empresas segundo o SEBRAE

Classificação	Indústria	Comércio e Serviço
Microempresa	Até 19 empregados	Até nove empregados
Pequena	De 20 a 99 empregados	De 10 a 49 empregados
Média	De 100 a 499 empregados	De 50 a 99 empregados
Grande	Mais de 500 empregados	Mais de 100 empregados

Fonte: Elaborado pelo autor

No estado de São Paulo existem 1.118.986 pequenos negócios empresariais de serviços, o que representa 41% do total de pequenos negócios do Estado (SEBRAE-SP, 2017). Por segmentos de atividade, destacam-se restaurantes e outros estabelecimentos de serviços de alimentação e bebidas (13,9% dos pequenos negócios de serviços), cabeleireiros e outras atividades de tratamento de beleza (13,1%) e transporte rodoviário de carga (6,8%) (SEBRAE-SP, 2017).

Entretanto, apesar de sua importância, muitas destas empresas tendem à certo grau de informalidade (NUNES *et al.*, 2006) e possuem baixo apreço às práticas de gestão da informação e do conhecimento em função da carência de recursos, muitas vezes até

desconhecendo seus benefícios. Isto se deve, sobretudo, a algumas características comuns neste tipo de empresa. Nas PMEs, muitas vezes o proprietário do negócio e poucos funcionários desempenham várias atribuições estratégicas, tais como vendas e atendimento. Isto significa que a falta ou a saída de qualquer um desses profissionais pode implicar numa grande perda ou até mesmo o fim da empresa (DURST; WILHELM, 2013).

Outra questão importante volta-se ao fato de que, normalmente, as pequenas e médias empresas não possuem recursos para investir em terra, trabalho e capital (DESOUZA; AWAZU, 2006). Em função disso, elas são forçadas a realizar mais com menos recursos. Nesse contexto, a utilização sistemática do conhecimento das pessoas envolvidas no negócio e, sobretudo, de entidades externas, tais como os clientes, pode ser uma alternativa interessante para obtenção de vantagem competitiva. Isto porque, conforme apregoam Nonaka e Takeuchi (2009), o conhecimento se multiplica quando é compartilhado, ao contrário do que acontece com terra, capital e trabalho.

Em complemento ao conhecimento como importante recurso, as tecnologias da Web 2.0, sobretudo as redes sociais, proporcionam formas dinâmicas de interação entre pessoas e negócios (O'REILLY, 2005). A importância e o impacto que estas tecnologias exercem sobre pequenos e médios negócios vai muito além da simples presença online e da facilidade de comunicação com seus clientes (SCHIVINSKI; DABROWSKI, 2016).

Ou seja, as redes sociais fornecem recursos específicos para explorar a interatividade empresa-cliente a partir de uma perspectiva baseada na percepção mútua destes elementos. Em primeiro lugar, as redes sociais incorporam recursos de mídia de massa tradicional e comunicação interpessoal em uma única plataforma (BOYD, 2008; MARWICK; BOYD, 2011; VITAK, 2012). Assim, empresas e consumidores podem se comunicar de forma transparente entre si por meio de redes sociais. Em segundo lugar, a visibilidade é alta porque as redes sociais permitem que as entidades de massa (empresas) e os indivíduos (consumidores) tornem seus conteúdos observáveis para uma ampla audiência (BOYD, 2010; TREEM; LEONARDI, 2012; MARWICK, 2015).

Além disso, estudos indicam que há forte relação entre a presença social das empresas percebida pelos clientes, a confiança na marca e a lealdade destes em relação a estas empresas, impactando diretamente em fatores tais como o engajamento e lealdade destes clientes em relação às marcas (HOLLEBEEK, 2011; ENGINKAYA; YILMAZ, 2014; HOLLEBEEK; GLYNN; BRODIE, 2014; PONGPAEW; SPEECE; TIANGSOONGNERN, 2017).

Além da grande visibilidade e possibilidade de atingir um contingente enorme de clientes em potencial, um fator determinante para o uso de redes sociais por parte de pequenas e médias empresas é o binômio 'custo x impacto'. Assim, as empresas são particularmente atraídas pelo baixo custo, quantidade crescente de assinantes e forte interatividade do marketing aplicado em redes sociais, adotando-se para tanto, as mídias sociais em seu composto de comunicação para com o cliente (MICHAELIDOU; SIAMAGKA; CHRISTODOULIDES, 2011; SMITS; MOGOS, 2014; ZEMBIK, 2014).

No entanto, a mera presença de uma empresa em mídias sociais não garante interações vantajosas entre empresa e consumidores (VENDEMIA, 2017). É necessário que a empresa que adota estas tecnologias esteja de fato preparada para lidar com os clientes por meio de redes sociais.

As redes sociais configuram-se no canal ideal para manter relações proveitosas, tanto para clientes, quanto para empresas dado que, conforme afirmam Bhattacharya e Sem (2003), interações bem sucedidas entre empresa e consumidor promovem a fidelidade do cliente, a vontade de experimentar novas ofertas da empresa e a resistência à informação negativa sobre a empresa.

Face ao contexto exposto, esta pesquisa busca contribuir para o avanço da temática de gestão do conhecimento do cliente oriunda da mineração de textos em redes sociais.

1.2 Problema de Pesquisa

Baseado no contexto apresentado até então, apresenta-se a seguinte questão de pesquisa que norteará esta dissertação: **Como um *framework* para mineração de opiniões pode ser aplicado para a descoberta de conhecimento do cliente em relação às suas experiências em restaurantes, com base em dados extraídos de redes sociais e que seja aplicável à realidade de pequenas e médias empresas?**

1.3 Objetivos

O objetivo geral desta pesquisa é apresentar um *framework* para mineração de opiniões para descoberta de conhecimento do cliente referente às suas experiências em

restaurantes oriundas de redes sociais, que seja aplicável à realidade de pequenas e médias empresas.

Com o intuito de atingir o propósito estabelecido pela pesquisa, define-se a seguir os seguintes objetivos específicos:

- Selecionar a rede social que melhor atende o domínio estudado;
- Realizar a extração dos dados da rede social escolhida;
- Converter os dados extraídos para os formatos apropriados à análise, considerando-se cada processo sugerido no *framework*;
- Aplicar Análise de Sentimentos e Modelagem de Tópicos para a exposição de conhecimentos provenientes dos clientes;
- Gerar visualizações a partir da massa de dados processada por meio do *framework* elaborado;
- Analisar os resultados gerados pelo *framework* estabelecido, com vistas a evidenciar o conhecimento do cliente a partir da mineração de dados da rede social selecionada.

1.4 Justificativa da pesquisa

A importância de considerar o envolvimento do cliente para o desenvolvimento do negócio é indiscutível às empresas de sucesso (DESOUZA; AWAZU, 2005). Cada vez mais nota-se o esforço destas na busca por estratégias que objetivam não somente desenvolver e ampliar relacionamentos com seus clientes, mas também utilizar suas ideias afim de melhorar produtos e serviços (CHUA; BANERJEE, 2013).

A contribuição dos clientes pode ser evidenciada por meio de vários aspectos que afetam diretamente a performance das empresas (KUMAR *et al.*, 2010). Afinal, atualmente os produtos e serviços fornecidos pelas empresas estão cada vez mais similares, sendo que a preferência de um cliente por uma ou outra marca se dá por pequenos detalhes (DAVENPORT; HARRIS, 2007). Assim, o conhecimento acerca do cliente pode viabilizar diferenciais em produtos e serviços, fazendo com que a empresa possa entregar um conjunto de valor mais apropriado aos diferentes perfis de cliente.

Os detalhes que regem estas escolhas são, na maior parte das vezes, subjetivos e, portanto, parte do arcabouço de conhecimento tácito que o cliente detém (BLACKWELL;

MINIARD; ENGEL, 2006). Alcançar o conhecimento tácito dos clientes é benéfico e essencial (NONAKA, 2006), sobretudo como forma de garantir a sobrevivência das pequenas e médias empresas, dado que este tipo de organização costuma não ter o mesmo nível de reservas para investimento em terra, trabalho e capital que as grandes empresas dispõem (DESOUZA; AWAZU, 2005).

A literatura sobre a gestão do conhecimento do cliente (*CKM – Customer Knowledge Management*) também evidencia a importância do conhecimento proveniente dos clientes como forma de melhorar o relacionamento das organizações com seus consumidores (SALOMANN *et al.*, 2005), além de enfatizar a relevância das interações com o consumidor como fonte vital de conhecimento para a empresa (GARCÍA-MURILLO; ANNABI, 2002; GEBERT *et al.*, 2003).

Segundo García-Murillo e Annabi (2002), a gestão eficiente do conhecimento oriunda dos clientes pode configurar-se numa fonte de vantagem competitiva, ao mesmo tempo em que tal interação ajuda as empresas a entenderem melhor seus clientes, proporcionando assim com que estas aprendam o que eles querem.

Sendo assim, a fim de obter vantagem competitiva e proporcionar valor adequado a ser entregue ao cliente, é importante que as empresas sejam capazes de alavancar seus ativos de conhecimento relacionados ao cliente, ativos estes que se constituem do conhecimento sobre e proveniente de seus clientes (DESOUZA; AWAZU, 2005). Como consequência direta deste processo, a distância entre a empresa e seu do cliente é estreitada, possibilitando à organização fornecer e desenvolver produtos e serviços que agreguem valor e que sejam orientados aos seus clientes (JAYACHANDRAN *et al.*, 2005).

Neste contexto, as ferramentas da Web 2.0 cumprem um papel crucial, uma vez que as empresas passaram a depender cada vez mais das mídias sociais para interagir com seus clientes (CHOUDHURY; HARRIGAN, 2014). As mídias sociais são definidas como um grupo de aplicativos baseados na Internet, que se fundamentam nas bases ideológicas e tecnológicas da Web 2.0 e que permitem a criação e troca de conteúdo gerado pelo usuário (KAPLAN; HAENLEIN, 2010).

Assim, as tecnologias da Web 2.0, sobretudo as redes sociais, proporcionam formas dinâmicas de interação entre pessoas e negócios (O'REILLY, 2005), ocasionando a melhoria da troca de informações entre a empresa e seus clientes, além de configurarem-se numa fonte de informações relevantes para as empresas (LEVY, 2009).

As redes sociais proporcionam às empresas a possibilidade de comunicação a custos irrisórios e de forma segmentada, criando assim, cada vez mais adeptos das corporações

(BLANCHARD, 2011; STELZNER, 2016). Outro aspecto fundamental das redes sociais para as organizações atuais reside em uma de suas principais características: o *feedback* quase instantâneo por parte dos clientes (BLANCHARD, 2011).

Embora a quantidade de opiniões disponível em redes sociais aumente a cada dia, a tarefa de analisar estes dados não é banal e, considerando a quantidade de informações a serem analisadas, não se trata de uma tarefa que possa ser desenvolvida de forma manual pelas empresas. Assim, a oportunidade (e ao mesmo tempo desafio) que se apresenta é a de conciliar a mineração de textos com o objetivo de revelar conhecimento a partir das opiniões dos usuários de rede sociais, sobretudo opiniões relacionadas às suas experiências com restaurantes, domínio escolhido para o desenvolvimento desta pesquisa de dissertação.

Dados existem em diversos formatos, de formas mais ou menos estruturadas, e nem sempre com características bem definidas, organizadas ou sequer disponibilizados em tabelas ou bancos de dados (BAARS; KEMPER, 2008). Na verdade, estima-se que a imensa maioria dos dados disponíveis atualmente sejam não-estruturados (FELDMAN; SANGER, 2007). Em 2020, existirão no mundo 50 vezes mais dados do que existiam em 2010, 70% a 80% destes constituindo-se de dados não-estruturados (HOLZINGER *et al.*, 2013), tais como textos. Devido ao volume e à variedade de dados não-estruturados, faz-se necessário desenvolver pesquisas que permitam avançar o conhecimento nessa área de investigação.

Conforme exposto, considerando-se a importância que as opiniões de clientes exercem sobre os negócios, bem como a ampla disponibilidade que as redes sociais oferecem no sentido de acesso direto às empresas, esta pesquisa propõe uma forma de extrair, processar e visualizar conhecimento a partir de opiniões de clientes provenientes de redes sociais por meio do emprego de técnicas de mineração de textos, análise de sentimentos e modelagem de tópicos.

1.5 Delimitações da pesquisa

Para tornar esta pesquisa viável, algumas considerações e, principalmente, algumas suposições foram determinadas. Uma suposição é qualquer ideia concebida que o pesquisador acredite ser verdade (GORDON; PATTERSON, 2013).

Assim sendo, inicia-se a delimitação do escopo desta pesquisa considerando um dos principais conceitos encontrados na literatura que versa sobre o Conhecimento do Cliente.

Há uma clara distinção em relação ao conhecimento sobre o cliente, conhecimento para o cliente e conhecimento do cliente (GEBERT et al., 2003). Esta pesquisa visa apresentar uma solução para extrair conhecimento do cliente, mas também considera que haja contribuição em relação ao conhecimento sobre o cliente, muito embora este segundo possa ser considerado mais um efeito colateral do que um objetivo em si. Portanto, é importante frisar que esta pesquisa não se propõe a contribuir em relação ao conhecimento para o cliente, dado que este conhecimento se encontra fora do escopo deste trabalho e domínio do pesquisador.

Outra delimitação diz respeito ao uso de dois termos que ocorrerão corriqueiramente neste trabalho: consumidor e cliente. Considerando que imensa maioria da literatura sobre o Conhecimento do Cliente é publicada em inglês e citada neste trabalho por meio de traduções e interpretações dadas pelo pesquisador, convém reforçar que ambos os termos não carregam em si nenhuma distinção. Muita embora o pesquisador tenha plena ciência do fato de que "a tradução de qualquer trabalho pode levar a uma perda de significado" (VAN NES *et al*., 2010), a literatura que fundamenta este trabalho não possui nenhuma definições que seja capaz de distinguir *'client'* (cliente) e *'customer'* (consumidor), inclusive empregando ambos os termos de forma intercambiável. Portanto, esta pesquisa toma como pressuposto que ambos os termos significam a mesma coisa.

Ainda em relação à adoção de expressões proveniente do idioma inglês, esta pesquisa adota de forma frequente os termos Redes Sociais (*Social Network*) e Mídias Sociais (*Social Media*). Muito embora os termos Mídia e Rede signifiquem, em si, coisas muito distintas, ambos são aplicados na literatura com o mesmo intuito: referir-se às formas de comunicação eletrônica (como sites de redes sociais e micro blogs) baseados em tecnologias da Web 2.0, através dos quais os usuários criam comunidades on-line para compartilhar informações, ideias, mensagens pessoais, opiniões e outros conteúdos, como fotos e vídeos (KIETZMANN *et al.*, 2011; LEONARDI, 2015). Portanto, assume-se, para os fins desta pesquisa, que os termos Redes Sociais e Mídias Sociais partilham do mesmo significado.

Outra definição importante refere-se ao emprego dos termos Mineração de Dados, Mineração de Textos e Mineração de Opiniões. As definições para cada um são claras na literatura que aborda cada uma das áreas.

Até recentemente, os especialistas de TI no mundo dos dados empresariais concentraram-se na "mineração de dados", que pode ser definida como a descoberta de conhecimento a partir de dados estruturados (dados contidos em bancos de dados estruturados ou *data warehouses*) (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Atualmente, a maioria dos dados comerciais disponíveis consiste em dados não estruturadas ou

semiestruturadas (FELDMAN; SANGER, 2007). Muito embora também possam conter números, datas e fatos em campos estruturados, a informação não estruturada consistem, em sua grande maioria, em textos (artigos, conteúdos textuais de site, postagens em blogs, conteúdos de mídias sociais, etc.). A presença de dados não estruturadas torna mais difícil realizar efetivamente atividades de gestão de conhecimento usando ferramentas tradicionais de inteligência de negócios.

Neste ponto reside a diferença crucial entre as áreas de Mineração de Dados e Mineração de Textos, dado que a primeira é voltada ao tratamento de dados estruturados e a segunda, dedica-se essencialmente a dados não estruturados, dentro os quais situam-se os conteúdos gerados por usuários de redes sociais, caso abordado nesta pesquisa, e que nos conduz à próxima definição: Mineração de Opiniões.

A área de Mineração de Opiniões, também conhecida como Análise de Sentimento ou Análise de Subjetividade (LIU, 2012; PANG; LEE, 2008) é uma disciplina que abrange pesquisas em áreas como mineração de dados, linguística computacional, recuperação de informações, inteligência artificial, entre outras. A Mineração de Opiniões é definida em (LIU, 2010) como qualquer estudo que intencione desenvolver soluções computacionais para o tratamento de opiniões, avaliações, impressões, sentimentos, afeições, visões, emoções e subjetividade, expressos em forma de conteúdo textual.

Para fins de delimitação e esclarecimento, esta pesquisa debruça-se sobre o problema de revelar conhecimento a partir de opiniões de clientes extraídas de uma rede social, opiniões estas constituídas de textos, ou seja, dados não estruturados, por meio do emprego de técnicas de Mineração de Textos, área que se utiliza de princípios e ferramentas oriundas da Mineração de Dados. Para os fins desta pesquisa, muito embora haja certa tangência entre os termos, cada um possui um significado próprio e esta pesquisa não ignora tais definições.

Por fim, uma última, mas não menos importante delimitação desta pesquisa refere-se à natureza das opiniões de usuários de redes sociais. Há diversos estudos na literatura que abordam os motivos pelos quais as pessoas se dedicam ao uso de determinadas tecnologias, como por exemplo as redes sociais (DESSART; VELOUTSOU; MORGAN-THOMAS, 2015; LEONARDI, 2014, 2015; SCHIVINSKI; DABROWSKI, 2016; WANG; LEE; HUA, 2015).

Muito embora a rede social abordada nesta pesquisa tenha características muito claras e dedique-se a receber opiniões de pessoas sobre suas experiências em estabelecimentos, o que motiva alguém a opinar de forma honesta sobre estas experiências é algo que está fora do escopo desta pesquisa. Em outras palavras, este trabalho parte do pressuposto *naïf* que as opiniões que as pessoas criaram é autêntica e refletem, de fato, suas visões. Assim sendo, esta

pesquisa parte do pressuposto que as opiniões contidas na rede social escolhida são o mais honestas possível.

Outra consideração importante que convém abordar tange a origem dos dados utilizados nesta pesquisa, preocupação que tem, por sinal, um caráter ético. Gordon e Patterson (2013) observaram que preocupações éticas são uma parcela significativa das pesquisas qualitativas.

Os dados utilizados nesta pesquisa de dissertação foram extraídos de uma rede social e estão disponíveis publicamente para consulta. No ato de sua extração todo e qualquer campo que pudesse identificar os autores das opiniões foram suprimidos, dado que esta pesquisa não intenta revelar ou ressaltar qualquer visão pessoal de quem quer que seja. Esta, por sinal, é uma característica da área de Mineração de Opiniões: avaliar uma massa de dados em detrimento de opiniões particulares dos indivíduos. Para tanto, as opiniões foram tratadas de tal forma a ignorar qualquer informação que possibilitasse identificar os indivíduos que as geraram, reforçando assim, o caráter ético desta pesquisa.

1.6 Organização da Dissertação

Este trabalho foi organizado em cinco capítulos. O Capítulo 1 introduziu o tema da pesquisa, o problema de pesquisa, seus objetivos, a justificativa para sua realização e suas delimitações.

O Capítulo 2 foi dedicado à fundamentação teórica que sustentou a pesquisa. Esta, por sua vez, foi dividida em subcapítulos, quais sejam: a) Conhecimento do cliente, b) Mídias Sociais, c) Mineração de textos, d) Análise de Sentimento e, por fim; e) Modelagem de Tópicos.

No capítulo 3 são expostos o método e instrumentos de pesquisa, bem como o *framework* para mineração de opiniões, e trata dos procedimentos metodológicos adotados nesta pesquisa. Neste capítulo, aborda-se a classificação da pesquisa, apresenta-se o *framework* proposto para a mineração de opiniões, a abordagem *tidy data* para mineração de textos e as ferramentas empregadas, bem como os detalhes referentes à coleta e pré-processamento dos dados.

Optou-se por uma forma não ortodoxa para apresentar os detalhes operacionais desta pesquisa, aplicando-se cada um dos processos e analisando-se os resultados por meio de um experimento preliminar. Desta forma, é possível acompanhar cada passo da metodologia de

pesquisa, observando-se inclusive os resultados pretendidos quando de sua aplicação aos dados das empresas nos quais o *framework* foi experimentado.

O Capítulo 4 é dedicado à Apresentação e Análise dos Resultados, no qual aplica-se o *framework* aos dados de três empresas, buscando-se por fim, a consolidação do método num último tópico que expõe as observações, problemas e descobertas encontradas no decorrer das análises.

Por fim, o Capítulo 5 apresenta a conclusão da pesquisa, suas contribuições para a Academia e praticantes nas empresas, bem como suas limitações e sugestões de trabalhos futuros que os resultados desta pesquisa podem suscitar.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são apresentados os principais conceitos teóricos existentes na literatura da temática explorada e que serão empregados no âmbito desta pesquisa. O capítulo inicia tratando sobre Conhecimento, e mais especificamente, o Conhecimento do Cliente. Em seguida, dedica-se ao tema Mídias Sociais, começando pelo conceito de Web 2.0, que o origina. Logo após, apresenta os polos teóricos mais técnicos como Mineração de Textos, Análise de Sentimentos e Modelagem de Tópicos.

2.1 Conhecimento do Cliente

O conhecimento tornou-se o ativo mais importante e indispensável para qualquer empresa que busque sobreviver em um mundo cada vez mais conectado, onde produtos, serviços e valores circulam de forma nunca antes imaginadas, sendo assim mais valioso e poderoso que qualquer outro ativo físico ou financeiro (STEWART, 1997). Considerando a importância do conceito, faz-se necessário explorar algumas de suas principais definições, bem como sua aplicabilidade no contexto das empresas.

Talvez por tratar-se de um conceito complexo, o conhecimento é definido de várias formas na literatura. Davenport e Prusak (2000) afirmam que conhecimento não é nem dado nem informação, entretanto, está relacionado com ambos, e a diferença entre estes dois termos é, geralmente, uma questão de grau. Os autores também afirmam que conhecimento, dado e informação não são conceitos intercambiáveis, e que sucesso e fracasso organizacionais dependem, muitas vezes, de saber qual deles você precisa, qual você tem e o que é possível fazer com cada um.

Uma das definições mais simples sobre o conceito de conhecimento é a definição do mesmo como o processo de transformação de dados em informação e informação em conhecimento (DAVENPORT; PRUSAK, 2000). Os autores definem que esta transformação se dá por meio de comparações, consequências, conexões e conversações. Colocando-se em outros termos, conhecimento pode ser interpretado como os meios pelos quais uma informação pode ser comparada a outra informação conhecida, o impacto desta informação com relação à

tomada de decisão, como conhecimentos relacionam-se com outros, e o que outras pessoas pensam sobre tal informação (DAVENPORT; PRUSAK, 2000, p. 6).

Conhecimento difere-se de informação por tratar-se de crenças e comprometimento, e tem a ver com ação (NONAKA; TAKEUCHI, 1995, p. 58). Entretanto, Nonaka e Takeuchi citam que conhecimento e informação partilham algo em comum: ambos devem ser imbuídos de sentido.

De Long e Fahey (2000) definem o conhecimento como "um produto da reflexão humana e da experiência". Nonaka (1994) define o conhecimento como "crença verdadeira justificada".

No que se refere à natureza do conhecimento, o mesmo pode ser dividido em duas grandes dimensões: tácito e explícito (POLANYI, 1966; NONAKA; TAKEUCHI, 1995; DALKIR; LIEBOWITZ, 2011).

Conhecimento explícito é o conhecimento que, de alguma forma foi capturado, armazenado e, posteriormente, disponibilizado para compartilhamento. Conhecimento explícito, por ser expresso em palavras, números, é comunicado e compartilhado em forma de dados brutos, fórmulas cientificas, processos codificados ou princípios universais (NONAKA; TAKEUCHI, 1995).

Conhecimento tácito é uma forma um pouco mais complexa de conhecimento (POLANYI, 1966). De acordo com Nonaka e Takeuchi (1995), conhecimento tácito é algo extremamente pessoal e, portanto, difícil de formalizar, tornando-o difícil de comunicar ou de compartilhar com os demais. Além disso, o conhecimento tácito está profundamente enraizado nas ações e experiências dos indivíduos, assim como nas ideias, valores, crenças e emoções que as pessoas adotam (NONAKA; TAKEUCHI, 1995).

Independente da visão sobre a definição do conhecimento, os autores concordam que o mesmo só é fecundo quando dividido. Takeuchi e Nonaka (2009) afirmam que o conhecimento se multiplica quando é compartilhado. Daí a importância de se pensar em formas de compartilha-lo no âmbito das organizações, sobretudo se considerarmos o momento atual, no qual o conhecimento é cada vez mais considerado não apenas importante, mas também vital para as empresas.

Peter Drucker é geralmente creditado como sendo o primeiro a popularizar o conceito de economia baseada no conhecimento.

Em sua obra *"The Age of Discontinuity: Guidelines to Our Changing Society"*, o autor sugere que nossas economias avançadas devem mudar para depender do trabalho do conhecimento e não da força industrial. Ele também prevê o fim da era da massa, cuja produção

é baseada no trabalho e o advento da era da informação baseada no conhecimento, e defende que o conhecimento não é apenas poder, mas também propriedade (DRUCKER, 1992).

Um aspecto importante desta transição reflete-se na maneira como as empresas lidam com fontes externas de conhecimento. Uma destas fontes, e talvez a mais importante de todas, é o cliente. O consumidor está cada vez mais próximo da empresa, e demonstra um comportamento mais ativo, conectado, informado, tem poderes e é atuante (PRAHALAD; RAMASWAMY, 2004).

Do ponto de vista do comportamento do consumidor, o conhecimento pode ser definido como "a informação armazenada na memória" (BLACKWELL; MINIARD; ENGEL, 2006).

O conhecimento é um fator importante nas decisões de compra dos consumidores, influenciando a busca de informações e a avaliação de produtos e serviços. Formas distintas de conhecimento do consumidor são comumente citadas na literatura (BRUCKS, 1985), sugerindo que a "expertise" do consumidor compreende duas dimensões descritas como conhecimento objetivo e subjetivo (ALBA; HUTCHINSON, 1987, 2000).

O conhecimento objetivo é a informação atual e precisa, armazenada pelos indivíduos em sua memória de longo prazo. Este tipo de conhecimento é baseado, em grande parte, na aprendizagem cognitiva juntamente com experiência credível com muitas ofertas e marcas dentro de uma categoria de produto (aprendizagem instrumental) (ALBA; HUTCHINSON, 1987).

O conhecimento subjetivo é o nível de conhecimento percebido pelo consumidor ou nível de conhecimento "auto avaliado", mais precisamente descrito como familiaridade de classe de produto. Os consumidores normalmente superestimam seus níveis de especialização, criando uma lacuna entre a percepção do que eles acreditam ser verdade sobre produtos e um julgamento mais preciso. Evidências empíricas estabeleceram que os consumidores, em sua maioria, não possuem o nível ou a qualidade de conhecimento objetivo que acreditam possuir (ALBA; HUTCHINSON, 1987, 1987; HEIMBACH; JOHANSSON; MACLACHLAN, 1989; ALBA; HUTCHINSON, 2000).

Sendo assim, não é surpreendente que muitos consumidores julguem mal a qualidade de produto através de pesquisas limitadas e interpretações errôneas dos atributos do produto, sejam eles intrínsecos ou extrínsecos.

A psicóloga do consumidor Merrie Brucks (1985) proporciona uma visão mais objetiva dessa diferença ao distinguir os dois tipos de conhecimento do consumidor:

- Conhecimento subjetivo - o que o consumidor acha que sabe;

- Conhecimento objetivo - o conhecimento real que um indivíduo possui e que pode ser medido por algum tipo de teste.

Nos últimos anos, o conhecimento do cliente teve seu valor reconhecido por estudiosos enquanto disciplina e, além disso, como uma fonte estratégica fundamental para o sucesso de qualquer empresa (ROWLEY, 2002; ROLLINS; HALINEN, 2005). Além disso, o conhecimento do cliente é apontado na literatura como forma de suportar o relacionamento de longo prazo com os clientes (DARROCH; MCNAUGHTON, 2003).

De acordo com Campbell (2003) o conhecimento do cliente refere-se ao entendimento das necessidades, expectativas e objetivos do cliente, e trata-se de um componente essencial para qualquer empresa que tenha a pretensão de construir relacionamentos reais com seus clientes.

Por outro lado, Paquette (2011) vai além, e sugere que o conhecimento do cliente pode ser composto por uma combinação de conhecimentos advindos de várias frentes, tais quais, o conhecimento do consumidor, conhecimento da cadeia de suprimentos, conhecimento específico da *joint venture* e assim por diante. Segundo autor, este conhecimento é proveniente de uma via de mão dupla, capaz de gerar valor para cliente e empresa. Vai muito além de informações que identificam os clientes, para um conhecimento que reside fora da organização.

Como exemplos disto, pode-se citar as preferências do consumidor para novos produtos, o conhecimento proveniente de pesquisas desenvolvidas em conjunto por empresas, melhorias de design propostas por fornecedores com o intuito de reduzir custos de manufatura e o conhecimento sobre tendências no ambiente de negócios, além de opiniões de clientes expressas em redes sociais ou outros canais (PAQUETTE, 2011).

Entretanto, um aspecto importante do conhecimento do consumidor é que tal conhecimento não pertence à empresa, mas sim, a terceiros que desejam ou não compartilha-lo. Ao mesmo tempo, com o alcance e profusão das tecnologias da Web 2.0, sobretudo as redes sociais (KAPLAN; HAENLEIN, 2010; KIETZMANN *et al.*, 2011; LEONARDI, 2014; LIN; LU, 2011; ZEMBIK, 2014), percebe-se uma predisposição muito grande dos usuários de mídias sociais ao compartilhamento de opiniões e ideias (LIN; LU, 2011).

O cliente pode fornecer conhecimento único que permite com que as empresas aprendam e melhorem suas operações internas (PAQUETTE, 2011). Além disso, a relação entre a empresa e o cliente é descrita como um processo dinâmico, onde ambos mudam ao longo do tempo (NEJATIAN *et al.*, 2011), o que, por sua vez, pode exige um esforço por parte da empresa não somente em se aproximar deste cliente, como também em tentar extrair deste suas opiniões e impressões sobre suas experiências de forma constante.

Entretanto esta não é uma tarefa banal, e exige não somente uma grande disposição por parte das empresas como também o emprego de tecnologias que permitam extrair e compreender o que está por trás destas opiniões.

A proposta deste trabalho é apresentar uma forma de extrair, processar e visualizar conhecimento a partir de opiniões de clientes provenientes de redes sociais por meio do emprego de técnicas de mineração de textos, análise de sentimentos e modelagem de tópicos.

Para isso, faz-se necessário compreender o conceito de mídias sociais, dado que esta é a origem dos dados tratados pelo *framework* proposto nesta pesquisa.

2.2 Mídias Sociais

A mídia social (ou rede social) é um dos termos mais populares da atualidade, sendo geralmente usado para referir-se a uma série de aplicações com características muito diferentes. O termo muitas vezes vem acompanhado de outros conceitos, razão pela qual a primeira parte deste tópico dedica-se a definir seus princípios e aplicações, bem como sua importância para Pequenas e Médias Empresas.

Considerando o objetivo desta pesquisa em viabilizar a captura de conhecimento do cliente a partir de opiniões extraídas de redes sociais, faz-se necessário elucidar o que vem a ser a Web 2.0, conjunto de princípios a partir do qual originam-se as redes sociais, causadoras, por sua vez, de mudanças na maneira como as pessoas compartilham conteúdos e opiniões on-line.

A Web 2.0 é um termo abrangente, cunhado inicialmente em 2005 por Tim O'Reilly (O'REILLY, 2005) e usado desde então para descrever uma diversidade de aplicações baseadas na web. A Web 2.0 não é uma nova versão da web, mas refere-se a novas formas de usar a internet para gerar conteúdo, explorar conexões entre usuários e encorajar a participação e transparência (O'REILLY, 2005, 2007).

Algumas aplicações e ferramentas comumente referenciados como exemplos da Web 2.0 incluem blogs, *podcasts*, *feeds* RSS, redes sociais, serviços de compartilhamento de fotos e vídeos, *wikis*, ambientes compartilháveis de conteúdos favoritos e serviços de distribuição de conteúdo *peer-to-peer* (O'REILLY, 2005).

De acordo com Kaletka e Pelka (2011), a Web 2.0 pode ser considerada uma das mais influentes e importantes inovações no campo da Tecnologia de Informação e

Comunicação, responsável por originar plataformas consagradas mundialmente como inovações em si, tais como Wikipedia, Facebook e YouTube.

Um erro comum e constante nessa temática é a confusão entre os conceitos de Web 2.0 e de mídias sociais. Há de se ressaltar que esses termos definitivamente não significam a mesma coisa, embora sejam empregados como sinônimos em algumas circunstâncias. Cormode e Krishnamurthy (2008) distinguem os dois conceitos afirmando que Web 2.0 é um espaço onde usuários individuais são tratados como objetos de primeira classe mas, ao mesmo tempo, trata-se de uma plataforma na qual diversas tecnologias inovadoras foram construídas ao longo dos anos, acomodando assim diferentes redes sociais modernas, tais como Facebook e Twitter, dentre outras.

Kaletka e Pelka (2011) argumentam que a inovação inicialmente introduzida pela Web 1.0 era baseada, sobretudo, na ideia de permitir que diferentes usuários interagissem com produtores e transmissores de conteúdo por meio de sites estáticos e sistemas de bate-papo e e-mails. Cormode and Krishnamurthy (2008) afirmam que boa parte dos usuários da Web 1.0 agia como simples consumidores de conteúdo, principalmente por não existirem muitos produtores de conteúdo no início. Para os autores, no caso da Web 2.0 praticamente todos os usuários também desempenham o papel de criadores de conteúdo, devido à imensa quantidade de ferramentas tecnológicas e plataformas de criação e compartilhamento disponíveis na Web 2.0.

John (2013) argumenta que o aspecto compartilhamento seja justamente a principal atividade da Web 2.0 e que as redes sociais se baseiem exatamente neste princípio. Como exemplos de plataformas nas quais o compartilhamento é definido como principal atividade, o autor cita Facebook, YouTube, Flickr, Twitter, wikis e blogs. Embora estas aplicações pareçam não ter relação direta entre si, elas compartilham algumas tecnologias em comum que as distinguem da tecnologia disponibilizada na Web 1.0, conforme exposto no Quadro 2.

Quadro 2 - Comparação entre as características da Web 1.0 e Web 2.0

Característica	Web 1.0	Web 2.0
Indexação e recuperação de informações	Classificação hierárquica (Taxonomias)	Classificação não-hierárquica (*Tags* ou *Folksonomies*)
Fluxo de informação	De cima para baixo	Horizontal ou inferior
O nível de interatividade	Mão única ou Mão dupla assimétrica	Mão dupla assimétrica ou simétrica
Papel dos usuários	Audiência	Participante
Papel do administrador	Publicador	Parceiro ou Protetor do conteúdo
O objetivo da comunicação	Entrega eficiente de informações	Entrega eficiente, entendimento mútuo
Tipo de comunicação	Modelo de publicação	Modelo de diálogo

Fonte: Elaborado pelo autor

Casey e Li (2012) denotam que, embora quase todas as tecnologias da Web 2.0 tenham como principal característica encorajar um certo nível de interação entre os usuários, as tecnologias apresentadas no Quadro 3 variam desde ferramentas extremamente interativas (o equivalente ao cara-a-cara ou reuniões em grupo) até tecnologias menos interativas, que permitem a comunicação ou troca de informações.

Quadro 3 - Tecnologias da Web 2.0 e seu potencial

Tecnologia	Potencial
Blogs	Fornecer informações para novas audiências usando um tom informal, possibilita conversas públicas, canal para resolver problemas.
Wikis	Colaboração em grupo de trabalho ou público para gerenciamento de projetos, compartilhamento de conhecimento e informações.
Compartilhamento de vídeo e multimídia	Apoio público, educação, treinamento, outra forma de comunicação para públicos "conectados" e on-line. Vídeos e áudios para melhorar o serviço e suporte, treinamento e educação de funcionários e gestores.
Compartilhamento de Fotos	Potencial de economia de custos, atração de novas audiências. Aplicações em marketing e divulgação de produtos e serviços.
Podcasting	Outra ferramenta para divulgar informações. Útil para construir confiança com a voz conversacional, atualizações de projetos, transmissões ao vivo ou mensagens de instruções.
Mundos Virtuais	Divulgação pública, prefeituras virtuais, educação, treinamento, capacidade de reunir pessoas em todo o mundo para reuniões, palestras, etc.
Redes Sociais	Impacto viral, gestão do conhecimento, recrutamento, anúncios de eventos e trocas de informações, publicação e compartilhamento de opiniões e outros tipos de conteúdos gerados pelo usuário.
Syndicated Web	Expandir o alcance e reunir conteúdo em coleções.
Feeds	Fonte autorizada, reduz a duplicação de informações e mantem as pessoas atualizadas.
Mashups	Expandir o alcance, prover serviço, integrar dados externos, disponibilizar conteúdos para outros que usam mashups, deliberação adotada e identificação de questões.
Widgets, Gadgets, Pipes	Aumentar a consciência do que está acontecendo, traz conteúdo e informações importantes para a página inicial do usuário.
Bookmark Social e Notícias	Aumentar a popularidade e o uso de sites específicos, em formação e serviços.
Micro-blogging	Transmissão de mensagens, anúncios e relatórios em tempo real.

Fonte: Adaptado de Casey e Li (2012).

Conforme indicam Cormode e Krishnamurthy (2008), ao contrário da World Wide Web estática, a Web 2.0 torna o usuário um objeto de primeira classe em seus sistemas e, além disso, torna a interação mais fácil para o usuário. Os autores apresentam algumas características fundamentais que definem as aplicações baseadas na Web 2.0:

- Usuários são considerados entidades de primeira classe no sistema, com páginas de perfil proeminentes, incluindo informações pessoais como idade, sexo, localização, depoimentos ou comentários sobre o usuário por outros usuários;
- Habilidade de formar conexões entre os usuários, por meio de conexões com outros usuários denominados ‘amigos’. Há ainda a adesão ou associação a grupos de vários tipos, bem como subscrições ou *feeds* RSS (*Rich Site Summary*) de atualizações de outros usuários;

- Capacidade de postar conteúdo de vários tipos (fotos, vídeos, blogs, comentários e classificações em conteúdo) gerados por outros usuários, além da possibilidade de adicionar tags em conteúdos próprios ou de outros usuários, bem como algum tipo de controle de privacidade e compartilhamento dos conteúdos.

Outro esforço no sentido de distinguir a Web 2.0 de outras tecnologias como a Web 1.0 pode ser encontrada na obra de Bradley (2009). O autor identifica seis características fundamentais da Web 2.0: participativa, coletiva, transparente, independente, persistente e emergente. Assim, a utilização da Web 2.0 é baseada na participação e esforço coletivo a qualquer momento e em qualquer lugar. Tal esforço ocorre num ambiente transparente que leva ao surgimento de ideias e conteúdos que permanecem persistentes para uso futuro (BRADLEY, 2009).

Nesse contexto, a característica social da Web 2.0 tem como principal expoente as tecnologias e plataformas conhecidas como mídias sociais ou redes sociais (LIN; LU, 2011). São aplicativos que possibilitam não apenas o consumo de conteúdo, mas também a produção e publicação destes, bem como a interconexão entre usuários por meio de diversos tipos de interações, sejam diretas (um para um) ou indiretas (um para muitos e muitos para muitos) (KIETZMANN *et al.*, 2011).

Alguns destes serviços online foram concebidos exclusivamente para a criação ou manutenção de relações sociais entre os usuários, enquanto outros têm como objetivo facilitar o compartilhamento de arquivos criados por usuários-membros (sites de compartilhamento de conteúdo gerado pelos usuários) (LIN; LU, 2011). A popularidade desses serviços tem oferecido uma variedade de possibilidades e oportunidades para os usuários, não só ao nível interpessoal e criativo, mas também em nível profissional. Portanto, o estudo destas redes sociais e das interações de seus usuários tornou-se uma seara interessante para a comunidade científica.

O conceito de mídias sociais é muito amplo e, por vezes, sem definição consensual na literatura. Um exemplo disso é a definição de Correa *et al.* (2010), que conceitua mídias sociais como uma forma de consumo específico de conteúdos digitais, embora o autor deixe claro que as mídias sociais tenham pouco a ver com o uso tradicional de mídias informativas em geral. Por outro lado, há definições mais amplas e até mesmo de maior complexidade, como é o caso da contribuição de Kaplan e Haenlein (2010), que definem mídias sociais como um

conjunto de aplicações para a Internet construídas com base nos fundamentos ideológicos e tecnológicos da Web 2.0, e que permitem a criação e troca de conteúdos gerados pelos usuários.

Embora as definições acima citem os usuários de mídias sociais ou redes sociais de forma bastante genérica, as mídias sociais não estão confinadas apenas ao uso pessoal. Assim, Kietzmann *et al.* (2011) citam o uso de mídias sociais por empresas que buscam com isso aumentar o retorno financeiro da companhia e melhorar a imagem da marca através de sua presença online. Os autores defendem a ideia de que tais tecnologias sejam, atualmente, fundamentais para a sobrevivência das empresas contemporâneas.

Uma outra abordagem para definir o papel das mídias sociais repousa na ideia de analisar suas *affordances*. O conceito de *affordance*, termo atualmente sem tradução para o português, pode ser entendido como 'reconhecimento de potencial', tendo a ver com a qualidade de um objeto que permita ao indivíduo identificar sua funcionalidade ou potencial, sem a necessidade de prévia explicação, o que ocorre de forma intuitiva (GIBSON, 1986; NORMAN, 2002). Dessa forma, o conceito foi originalmente criado por James J. Gibson, um psicólogo americano que desenvolveu inúmeros trabalhos no campo da percepção visual. Treem e Leonardi (2012) apresentam uma aplicação da teoria das *affordances* para explicar como a interação dos usuários de mídias sociais corporativas difere da interação com formas tradicionais de comunicação mediadas por computador.

Para fins desta dissertação serão consideradas as visões de Kietzmann *et al.* (2011) e Gibson (1994), que consideram que diferentes indivíduos adotam visões diferentes e particulares sobre como lidar com as mais diversas tecnologias.

Aborda-se a seguir o tópico sobre Mineração de Textos, que também versará sobre Processamento de Linguagem Natural e a Extração de Dados, conceitos fundamentais para o desenvolvimento do *framework* de mineração de opiniões proposto por esta pesquisa.

2.3 Mineração de Textos

A língua é a principal manifestação da inteligência humana (PINKER, 2008; GARDNER, 2011). Através da linguagem expressam-se desde necessidades básicas a conhecimentos técnicos. A linguagem tornou-se questão central da filosofia no século XX, dando origem a diversas abordagens, como por exemplo, a Hermenêutica, Fenomenologia, Filosofia Analítica e o Estruturalismo, para citar algumas. Uma abordagem mais recente ganhou

força graças às pesquisas de áreas como neurociências, genética comportamental e psicologia evolucionista, que empreendem esforços objetivando compreender como o homem capta, organiza e transforma informações e sinais do meio em conhecimento e o expressa por meio da linguagem e de comportamentos (PINKER, 1998, 2008).

Em sua obra *"Do que é feito o pensamento"*, o psicólogo evolucionista e linguista canadense Steven Pinker fala, entre outras coisas, sobre o papel da linguagem, que define como uma das maravilhas do mundo natural. A capacidade de nos comunicar de forma compreensível com nossos pares demonstra uma característica natural da espécie humana (PINKER, 2008).

Segundo o autor, a linguagem é produto de um processo evolutivo da mente humana que foi se adaptando ao longo dos tempos para enfrentar e superar condições adversas visando a própria sobrevivência, e está diretamente ligada a um processo cognitivo vinculado diretamente ao pensamento e aos múltiplos contextos que formam o meio (PINKER, 2008).

Entretanto, a ideia da linguagem como instinto não é de Pinker. Ela é fortemente influenciada pelas ideais de Charles Darwin e do linguista Noam Chomsky, que revolucionou a ciência cognitiva ao afirmar a existência de uma gramática mental no nosso cérebro, que possibilita a construção de uma infinidade de frases a partir de combinações entre uma quantidade determinada de palavras (CHOMSKY, 2009).

Pinker defende a ideia da linguagem como instinto para ilustrar algo inerente à linguagem humana, também conhecida como linguagem natural: sua complexidade e mutabilidade. Segundo o autor, linguagem natural evolui e é reinventada de geração em geração, e embora tenha regras, estas são pouco estáticas (PINKER, 2002), ao contrário do que acontece com linguagens artificiais, como as de programação, que são extremamente precisas, contendo regras e estruturas lógicas fixas e bem definidas, permitindo aos computadores saberem exatamente como proceder a cada comando.

No entanto, em se tratando de linguagens humanas, uma simples frase pode conter ambiguidades, nuances e inflexões interpretativas que dependem de um determinado contexto, do conhecimento do mundo, de regras gramaticais e culturais e de conceitos abstratos (JACKSON; MOULINIER, 2007).

Quando as pessoas se comunicam, o fazem de muitas maneiras: escrevem livros, artigos, blogs e páginas da web, interagem enviando mensagens de diferentes maneiras e, claro, falam umas com as outras. Quando isto acontece eletronicamente, esses dados de texto tornam-se um recurso significativo, que tem enorme valor potencial para uma ampla gama de organizações (FELDMAN; SANGER, 2007).

Devido sobretudo ao avanço das tecnologias da Web 2.0 o volume de dados e a velocidade com que podem ser consultados são mais elevados a cada dia que passa. No entanto, a capacidade das pessoas para processar e compreender esses dados permanece constante (HOFMANN; CHISHOLM, 2013). Para superar a limitação humana, a área de Processamento de Linguagem Natural (PLN) dedica-se a investigar, propor e desenvolver sistemas computacionais que têm a linguagem natural escrita como objeto de estudo (GRISHMAN, 1986).

De uma forma mais simples, o Processamento de Linguagem Natural pode ser definido num sentido amplo como qualquer tipo de manipulação da linguagem natural em computador, o que pode envolver diversos métodos e técnicas (CLARK; FOX; LAPPIN, 2010).

Liddy (2001) define o Processamento de Linguagem Natural como um conjunto de técnicas computacionais motivadas pela análise e representação de textos que ocorrem naturalmente em um ou mais níveis de análise linguística para fins de processamento de linguagem humana e envolve uma variedade de tarefas e aplicações.

Desta definição, a autora destaca a noção imprecisa de conjunto de técnicas computacionais, pois existem vários métodos ou técnicas a serem escolhidas. Ela também frisa que os textos de ocorrência natural podem ser em qualquer idioma, modo ou gênero, orais ou escritos. A única exigência é que eles sejam de uma linguagem utilizada pelo homem para se comunicar entre si (LIDDY, 2001).

A autora considera ainda que existem vários níveis de análise linguística com diferentes significados e que os sistemas de Processamento de Linguagem Natural utilizam diferentes níveis ou combinações desses níveis. Sendo assim, o Processamento de Linguagem Natural torna possível a computadores compreender e interpretar as mais diversas instruções fornecidas por uma determinada linguagem natural (LIDDY, 2001).

O processo de extração de valor de dados de texto, conhecido como Mineração de Texto, é uma das áreas de conhecimento que mais ganhou atenção dos profissionais de mercado e de pesquisadores da academia nos últimos tempos, sobretudo graças à evolução da internet e dos meios de comunicação móveis (BERRY; KOGAN, 2010).

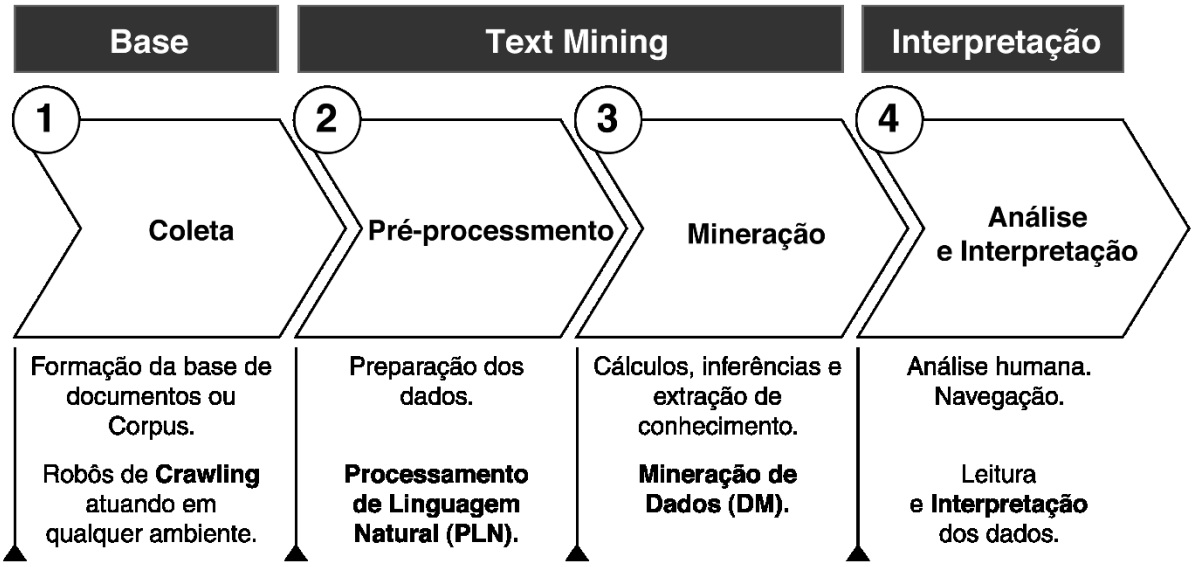
Em função disso, diversas técnicas e processos foram desenvolvidas com o intuito de recuperar informações relevantes contidas em bases de dados não-estruturados, dando origem à área conhecida como Mineração de Textos (do inglês *Text Mining*) que, por sua vez, deriva de uma área mais ampla conhecida como Mineração de Dados (do inglês *Data Mining*) (FELDMAN; SANGER, 2007).

A Mineração de Textos é essencialmente uma atividade multidisciplinar e, além de empregar certas técnicas da Mineração de Dados, apoia-se também em outras áreas, tais como a Inteligência Computacional, Recuperação da Informação, Ciência Cognitiva e, sobretudo, o Processamento de Linguagem Natural (CHOWDHURY, 2003).

Consta na literatura uma série de abordagens em relação aos processos de Mineração de Textos. Entretanto, estes processos podem ser resumido em quatro grandes etapas: (1) extração de dados (também referenciada como coleta de documentos); (2) pré-processamento; (3) extração de padrões; e (4) análise e avaliação de resultados (FELDMAN; SANGER, 2007; ARANHA, 2007).

Aranha (2007) apresenta um modelo completo para aquisição de conhecimentos a partir de textos, definindo inclusive, a distinção de etapas a serem desempenhadas por técnicas computacionais e por pessoas, bem como a consideração da etapa inicial de coleta de dados como parte da metodologia, conforme exposto na Figura 1.

Figura 1 – Metodologia de Mineração de Textos proposta por Aranha



Fonte: Adaptado de Aranha (2007).

A seguir, descreve-se resumidamente cada uma das fases da metodologia proposta por Aranha (2007).

A etapa de **Coleta**, fase inicial do processo, tem como objetivo formar uma base de dados textuais, conhecida na literatura como *corpus* (SILVA; PERES; BOSCARIOLI, 2017). Segundo os autores, esta fase pode ocorrer de várias maneiras, porém, seja qual for a forma

adotada, todas necessitam de grande esforço, dada a dificuldade de se obter material de qualidade e que sirva de matéria-prima para a aquisição de conhecimento.

O **Pré-processamento** é a etapa executada logo após a Coleta e tem como objetivo tornar a massa textual mais uniforme. Segundo Silva, Peres e Boscarioli (2017), trata-se de uma atividade bastante onerosa, tanto em tempo quanto em processamento, requerendo, não raro, a aplicação de diversos algoritmos que consomem boa parte do tempo do processo de extração de conhecimento.

Uma vez obtidas a estrutura para os dados e a forma de indexação destes, a etapa de **Mineração** é iniciada. Nesta fase desenvolvem-se cálculos, inferências e aplicam-se algoritmos que tem como objetivo a extração de conhecimento e descoberta de padrões. Como consequência, verifica-se ao término desta etapa o surgimento de novos aspectos até então não conhecidos.

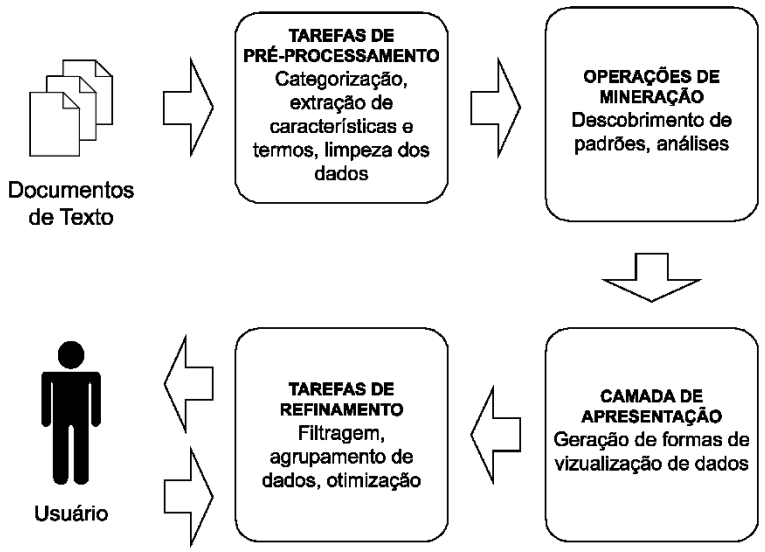
Por último, a **Análise** é a etapa que, segundo Aranha (2007), deve ser executada por pessoas que, de alguma forma, estão interessadas no conhecimento extraído. O autor ressalta que algum conhecimento acerca do domínio sobre o qual se desenvolve a mineração pode ser importante para a tomada de decisão apoiada no processo de Mineração de Textos.

De forma semelhante à proposta de Aranha (2007), Feldman e Sanger (2007), afirmam que, em um nível funcional, os sistemas de mineração de texto seguem o modelo geral fornecido por algumas aplicações clássicas de mineração de dados e, portanto, são segregados em quatro áreas principais:

a) Tarefas de pré-processamento;
b) Operações de mineração;
c) Camada de apresentação;
d) Tarefas de refinamento.

A Figura 2 mostra o processo de mineração de textos descrito por Feldman e Sanger (2007):

Figura 2 – Processo de Mineração de Textos



Fonte: Adaptado de Feldman e Sanger (2007).

As etapas para mineração de textos propostas por Feldman e Sanger (2007) são melhor descritas a seguir.

As tarefas de pré-processamento consistem nas rotinas, processos e métodos necessários para preparar os dados para o próximo estágio. Seu objetivo é formatar os dados originais de forma a torná-los acessíveis aos métodos de mineração. Em alguns casos é possível executar tarefas capazes de extrair ou aplicar regras para facilitar a formatação.

As operações de mineração correspondem ao núcleo de qualquer sistema de mineração de textos. Dentre as tarefas executadas nesse estágio, pode-se destacar o descobrimento de padrões, análises de tendência e algoritmos de descoberta de conhecimento. Os padrões mais utilizados para a descoberta de conhecimento em textos baseiam-se em conceitos de associação, frequência e distribuição dos dados. Tais conceitos podem ser aplicados individualmente ou por meio de comparações entre si, visando a obtenção de melhores resultados.

A camada de apresentação compreende principalmente a geração de formas de visualização de resultados das operações realizadas na fase de mineração.

As tarefas de refinamento envolvem, em sua forma mais simples, métodos que filtram informações redundantes e agrupam dados estreitamente relacionados. Entretanto, estas tarefas podem crescer, em um determinado sistema de mineração de texto, de modo a representar um conjunto completo e abrangente de alternativas de supressão, ordenação, poda,

generalização e agrupamento dos dados voltado à otimização de descobertas. Essas técnicas também são comumente descritas como pós-processamento.

Para o contexto desta dissertação, o *framework* de mineração de opiniões proposto baseia-se nos trabalhos de Feldman e Sanger (2007) e de Aranha (2007). Esta escolha deve-se a dois motivos: primeiro por ambos possuírem caráter universal, ou seja, são muito parecidos com diversas outras propostas, que variam em relação ao escopo ou objetivo, podendo conter outras etapas ou processos mais específicos. O segundo motivo tem a ver com o fato de que nenhum dos dois seria capaz de contemplar de forma cabal os objetivos propostos por esta pesquisa.

Portanto, este trabalho sugere a adoção de partes dos dois modelos e a adição de processos específicos que visam atender os objetivos propostos nesta pesquisa. Como parte do *framework* proposto consiste justamente na extração de opiniões a partir de redes sociais, aborda-se a seguir a extração de dados, dada a sua importância para o escopo desta pesquisa.

A democratização da internet e, mais especificamente, a popularização das redes sociais ocasionou mudanças permanentes nos hábitos dos consumidores (O'REILLY, 2005; KAPLAN; HAENLEIN, 2010; KALETKA; PELKA, 2011). A incorporação da internet no cotidiano das pessoas produz um rastro de informações digitais, denominado na literatura como *digital footprint* (GARFINKEL; COX, 2009). Estas informações estão armazenadas em servidores, redes sociais, sites de busca, comentários deixados em blogs, cadastros em sites de vendas, históricos de busca por produtos, e inúmeras outras possibilidades.

Ao mesmo tempo em que a abundância de conteúdo proporciona grandes oportunidades para a descoberta de conhecimento, também traz consigo um problema de solução não muito simples, que antecede à fase de mineração de textos e, consequentemente, anterior à descoberta de conhecimento: a extração de conteúdo a partir de sites da internet (MALIK; RIZVI, 2011; VARGIU; URRU, 2012; DEVIKA; SURENDRAN, 2013; KAMANWAR; KALE, 2016), processo por meio do qual iniciam-se várias propostas de sistemas de mineração de textos comuns na literatura (FELDMAN; SANGER, 2007; BERRY; KOGAN, 2010; WICKHAM; GROLEMUND, 2016; WIEDEMANN, 2016).

O problema da extração de dados a partir de web sites é particularmente complexo devido à natureza da construção dos sites, que mudou muito nos últimos anos, incorporando novas tecnologias e formas de interação, o que se reflete diretamente em sua estrutura e construção (O'REILLY, 2007).

Sites são construídos em torno do conceito de interação humana, onde a informação geralmente é entregue de forma visualmente estruturada (NIELSEN; LORANGER, 2007;

KRUG, 2013). Embora a informação seja facilmente interpretada pela interface do usuário, os dados de texto relevantes são fornecidos em diferentes padrões, o que os torna pouco adequados a processos de extração automáticos (MUNZERT, 2015).

Embora haja quem defenda a aplicação de métodos manuais a exemplo do 'copiar-e-colar' como forma de extração dos dados de sites (VARGIU; URRU, 2012), não há sombra de dúvida que tais métodos significam a tradução da ineficiência, sobretudo em se tratando de enormes massas de dados e sites com centenas e até milhares de páginas.

Para estes casos, a extração automática não somente é uma opção mas também uma necessidade (DEVIKA; SURENDRAN, 2013), dado que seria humanamente impossível lidar com a tarefa da extração de grandes massas de dados de forma manual sem cometer-se erros, uma característica comum a processos que envolvem a intervenção humana.

A tarefa de extrair dados de web sites por meios automáticos ou semiautomáticos é conhecida como *web scraping* (MUNZERT, 2015), termo que não possui uma tradução definitiva para o português, mas pode ser compreendido como "raspagem da web". O sentido do termo é exatamente o que a tradução sugere: a tarefa de extrair e coletar os dados textuais relevantes das fontes, eliminando a desordem, como código relacionado à interface do usuário ou publicidade no site.

Segundo Devika e Surendran (2013) muitas aplicações relacionadas a diversas áreas de negócios apoiam-se na coleta de informações por meio da web, que por sua vez são fundamentais aos processos de tomada de decisão. A extração de dados em formato digital, sobretudo a informação proveniente de websites, tornou-se nos últimos anos uma das técnicas mais utilizadas na área de Big Data (MARRES; WELTEVREDE, 2013).

Sistemas de extração de dados da Web são uma ampla classe de aplicativos de software que visam extrair informações de fontes da internet (BAUMGARTNER; GATTERBAUER; GOTTLOB, 2009, 2009). Um sistema de extração de dados da Web geralmente interage com uma fonte da Web e extrai dados armazenados nele: por exemplo, se a fonte for uma página HTML, a informação extraída pode consistir em elementos estruturais da página, bem como no texto completo da página em si. Eventualmente, os dados extraídos podem ser pós-processados, convertidos no formato estruturado mais conveniente e armazenados para uso posterior.

Os sistemas de extração de dados da Web possuem uso extensivo em uma ampla gama de aplicações, incluindo a análise de documentos baseados em texto disponíveis para uma empresa (como e-mails, fóruns de suporte, documentação técnica e jurídica, etc.), Inteligência de Negócios (BAUMGARTNER *et al.*, 2005), monitoramento de plataformas sociais (GJOKA

et al., 2010; CATANESE *et al.*, 2011), Bioinformática (PLAKE *et al.*, 2006), e muitas outras áreas.

A importância dos sistemas de extração de dados da Web decorre principalmente do fato de que uma quantidade cada vez maior de informações é continuamente produzida, compartilhada e consumida on-line (O'REILLY, 2005; LEVY, 2009; KAPLAN; HAENLEIN, 2010; KIETZMANN *et al.*, 2011; WAGNER; VOLLMAR; WAGNER, 2014). Os sistemas de extração de dados da Web permitem coletar essa massa de informações com o mínimo de intervenção humana.

A disponibilidade e análise dos dados coletados é um requisito incontestável para compreender fenômenos sociais, científicos e econômicos complexos que geram a própria informação. Por exemplo, colecionar traços digitais produzidos por usuários de plataformas sociais da Web como Facebook, Twitter e YouTube é o passo-chave para entender, modelar e prever o comportamento humano(KLEINBERG, 2000; NEWMAN, 2003; CORREA; HINSLEY; ZÚÑIGA, 2010; BACKSTROM *et al.*, 2012).

Do ponto de vista comercial, a Web fornece uma riqueza de informações de domínio público, possibilitando com que uma empresa possa adquirir e analisar informações sobre a atividade de seus concorrentes. Este processo é conhecido como Inteligência Competitiva (ZANASI, 1998; CHEN; CHAU; ZENG, 2002) e é crucial identificar rapidamente as oportunidades oferecidas pelo mercado, antecipar as decisões estratégicas dos competidores, bem como aprender com suas falhas e sucessos.

O design e implementação de sistemas de extração de dados na Web foi discutido a partir de diferentes perspectivas e alavanca métodos científicos provenientes de várias disciplinas, incluindo Aprendizado de Máquina e Processamento de Linguagem Natural (MALIK; RIZVI, 2011; DEVIKA; SURENDRAN, 2013; KAMANWAR; KALE, 2016).

Na concepção de um sistema de extração de dados da Web, muitos fatores devem ser levados em consideração, alguns deles são independentes do domínio de aplicação. Por outro lado, outros fatores dependem fortemente dos recursos específicos do domínio da aplicação. Como consequência, algumas soluções tecnológicas que parecem ser úteis em alguns contextos de aplicação não são adequadas em outros (KAMANWAR; KALE, 2016).

Na sua formulação mais geral, o problema da extração de dados da Web é difícil porque é limitado por vários requisitos. Conforme descrito por (JACKSON; MOULINIER, 2007; BERRY; KOGAN, 2010; DEVIKA; SURENDRAN, 2013; MUNZERT, 2015; KAMANWAR; KALE, 2016), os principais desafios que se pode encontrar no projeto de um sistema de extração de dados da Web podem ser resumidos da seguinte forma:

- As técnicas de extração de dados da Web muitas vezes exigem a ajuda de especialistas humanos. Um primeiro desafio consiste em fornecer um alto grau de automação, reduzindo ao máximo interações humanas. Os comentários humanos, no entanto, podem desempenhar um papel importante no aumento do nível de precisão alcançado por um sistema de extração de dados na Web. Um desafio relacionado é, portanto, identificar um meio-termo razoável entre a necessidade de construir procedimentos altamente automatizados de extração de dados na Web e o requisito de alcançar um desempenho mais preciso.
- As técnicas de extração de dados da Web devem ser capazes de processar grandes volumes de dados em um tempo relativamente curto. Este requisito é particularmente rigoroso no campo da Inteligência Competitiva e de Negócios, dado que uma empresa precisa realizar análises sensíveis ao tempo das condições do mercado.
- Aplicações no campo da Web social ou, mais em geral, as que tratam de dados pessoais devem fornecer garantias de privacidade sólidas. Portanto, tentativas potenciais (mesmo que não intencionais) de violar a privacidade do usuário devem ser adequadamente identificadas e neutralizadas.
- Abordagens que dependem de Aprendizagem de Máquina muitas vezes exigem um conjunto de treinamento significativamente grande de páginas da Web rotuladas manualmente. Em geral, a tarefa de rotular páginas é muito cara e propensa a erros e, portanto, em muitos casos, não se trata de uma abordagem que valha a pena aplicar.
- Muitas vezes, uma ferramenta de extração de dados da Web precisa extrair dados rotineiramente de uma fonte de dados da Web que pode evoluir ao longo do tempo. As fontes da Web estão em constante evolução e as mudanças estruturais acontecem sem aviso prévio, portanto, são imprevisíveis. Eventualmente, em cenários reais, surge a necessidade de adaptar esses sistemas, que podem deixar de funcionar corretamente caso lhe falte flexibilidade para detectar e enfrentar modificações estruturais de fontes da Web relacionadas.

Conforme descrito por Munzert (2015), a tarefa de *web scraping* pode ser melhor detalhada em três tarefas:

1. Coletar arquivos HTML de uma fonte ou site;
2. Determinar os padrões estruturais por trás dos quais a informação textual é apresentada nestes arquivos;
3. Aplicar os padrões reconhecidos para extrair dados no formato de saída desejado.

Quanto ao aspecto técnico, Devika e Surendran (2013) afirmam que o *web scraping* funciona como uma espécie de engenharia reversa, ou seja, trata-se do processo inverso de criação de uma página.

Para Marres e Weltevrede (2013) o *web scraping* constitui-se de uma série de passos nos quais dados estruturados são extraídos de uma desordem informacional, por meio da criação de processos flexíveis, dado que a estrutura de cada web site pode variar drasticamente. Os autores também afirmam que técnicas de *web scraping* torna viável às pesquisas sociais trabalharem com grandes massas de dados gerados pelos próprios usuários, que ultimamente acumulam-se em plataformas online, por sinal, objeto principal desta pesquisa.

Esta pesquisa fez uso da técnica de *web scraping* para realizar a extração de opiniões de clientes da rede social TripAdvisor Brasil, escolhida para este trabalho. O detalhamento técnico bem como as decisões relacionadas ao processo de extração de dados é apresentado no Capítulo 3 - Procedimentos Metodológicos

Uma coleção enorme dos dados digitais está disposta em forma textual. Estes dados textuais estão apresentados em linguagem natural e em sua grande maioria, não-estruturados. Sendo assim, há dificuldade para efetuar a extração de regras a partir de dados não estruturados e, portanto, tais dados não podem ser usados para predição ou qualquer outra função útil. Este é um dos principais motivos que impulsionaram o desenvolvimento da área de Mineração de Textos nos últimos anos (HOFMANN; CHISHOLM, 2013).

Segundo Feldman e Sanger (2007), o objetivo tanto da mineração de dados, quanto da mineração de textos é extrair informações úteis a partir da identificação e exploração de padrões com o emprego de técnicas computacionais.

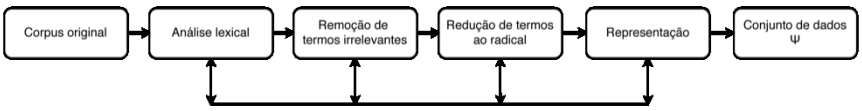
A diferença básica entre a mineração de dados e mineração de textos consiste no fato de que estes padrões não são encontrados em bases de dados estruturadas, mas sim em dados textuais não estruturados. Estes dados textuais podem ser completa ou parcialmente desestruturados (HAN; KAMBER, 2011).

Assim, para executar a mineração de textos é necessário tratar estes dados por meio de diferentes técnicas de refinamento aplicadas, não raro, sucessivamente. Estes dados refinados serão então organizados em estruturas mais apropriadas para possibilitar a extração de informações relevantes. Este processo é chamado de 'pré-processamento' (FELDMAN; SANGER, 2007; WEISS; INDURKHYA; ZHANG, 2010; ZHAI; MASSUNG, 2016).

Em mineração de dados o pré-processamento ocupa-se, basicamente, da normalização de dados já estruturados. Entretanto, em processos de mineração de textos, a etapa de pré-processamento tem o objetivo de identificar e extrair informações representativas de textos não estruturados. Esta importante etapa é responsável por transformar o texto desestruturado de uma base de documentos em um formato intermediário, mais estruturado, o que não é comum em sistemas de mineração de dados. Sendo assim, um sistema de mineração de textos adquire como entrada os documentos em forma de textos e gera diversos tipos de saída possíveis (FELDMAN; SANGER, 2007).

Uma outra abordagem em relação ao pré-processamento de dados pode ser encontrada em Silva, Peres e Boscarioli (2017), conforma mostra a Figura 3.

Figura 3 – Pré-processamento para obtenção do conjunto de dados Ψ



Fonte: Adaptado de Silva, Peres e Boscarioli (2017).

Segundo Silva, Peres e Boscarioli (2017) o processo exibido na Figura 3 apresenta uma sugestão de etapas para obtenção do conjunto de dados Ψ. Entretanto, as etapas escolhidas para o pré-processamento de dados devem considerar o domínio, área de atuação e os objetivos pretendidos (FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2017).

As decisões pertinentes ao pré-processamento das opiniões de clientes extraídas da rede social escolhida para esta pesquisa serão descritas com maior detalhamento no Capítulo 3 - Procedimentos Metodológicos.

O objetivo da etapa de extração de padrões é aplicar técnicas que possibilitem a extração de conhecimentos, utilizando-se, para tanto, de uma gama de algoritmos e técnicas de mineração provenientes de diversas áreas do conhecimento, tais como Aprendizado de Máquina, estatística e bancos de dados (ARANHA, 2007).

Entre as principais técnicas aplicadas à Mineração de Textos, pode-se destacar a Categorização, *Clustering*, Sumarização, Extração de Informação e Análise de Sentimento (NEVES; CORRÊA; CAVALCANTI, 2013).

Este trabalho emprega a Mineração de Textos com o objetivo de extrair conhecimento de opiniões de usuários publicadas em uma rede social em relação a restaurantes, domínio abordado nesta pesquisa. Para cumprir parte deste objetivo, propõe-se a utilização de Análise de Sentimentos, abordada a seguir.

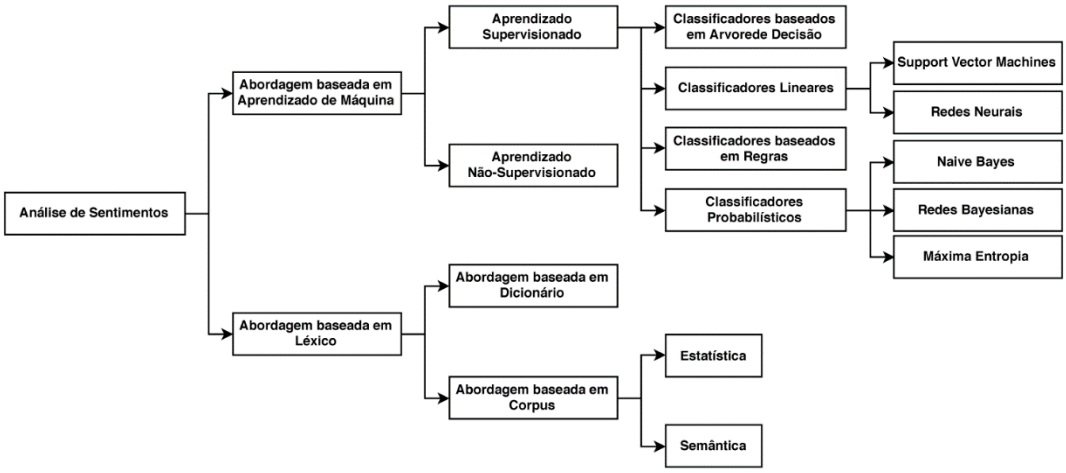
2.4 Análise de Sentimento

A Análise de Sentimentos, também conhecida como Mineração de Opinião, é uma área da Ciência da Computação que dedica-se ao problema de identificar emoções e opiniões em conteúdos textuais (PANG; LEE, 2004, 2008; LIU, 2012). Embora constate-se uma proliferação de estudos nesta área em anos mais recentes, o interesse por esse tema já existe há algum tempo (PANG; LEE, 2008).

A Análise de Sentimentos trata de problemas de classificação e, sendo assim, é utilizada para classificar textos de acordo como sua polaridade, seja ela positiva, negativa ou neutra. A Análise de Sentimentos é frequentemente abordada na literatura como uma atividade de classificação de sentimentos, ou ainda, classificação de polaridade de sentimentos, ou seja, positivo, negativo ou neutro (LIU, 2012). Assim, a Análise de Sentimentos está diretamente relacionada a um domínio específico, embora haja abordagens que tratem a questão do domínio de forma independente.

Walaa Medhat, Ahmed Hassan e Hoda Korashy (2014) realizaram extensa revisão de literatura sobre as diferentes aplicações e técnicas aplicadas à Análise de Sentimentos, conforme demonstra a Figura 4.

Figura 4 – Abordagens aplicadas na Análise de Sentimentos



Fonte: Adaptado de Medhat, Hassan e Korashy (2014).

A primeira abordagem utilizada na Análise de Sentimentos é o **aprendizado de máquina**, que inclui métodos de aprendizado automático não supervisionados e supervisionados.

De acordo com Mitchell (1997) o aprendizado de máquina estuda como os algoritmos computacionais são capazes de automaticamente melhorarem a execução de tarefas através da experiência. Os algoritmos de aprendizagem de máquina baseiam-se em estatística e probabilidade para aprender padrões complexos a partir de algum conjunto de dados.

A abordagem de aprendizagem não supervisionada usa conjuntos de dados não rotulados para descobrir a estrutura e encontrar padrões semelhantes a partir dos dados de entrada. O método não supervisionado geralmente é usado quando uma coleção de conjuntos de dados rotulados confiáveis é difícil, mas a coleta de dados não rotulados é mais fácil.

Turney (2002) usa abordagem de aprendizagem de máquinas não supervisionada para a classificação de avaliações. Os comentários são classificados como recomendados (*thumbs up*) e não recomendados (*thumbs down*). A classificação de uma revisão é prevista pela orientação semântica média das frases na revisão que contém adjetivos ou advérbios. Uma frase tem uma orientação semântica positiva quando possui boas associações (por exemplo, "nuances sutis") e uma orientação semântica negativa quando tem associações ruins (por exemplo, "muito cavalheiro").

Os métodos de aprendizagem de máquinas supervisionados assumem a presença de dados de treinamento rotulados que são utilizados no processo de aprendizagem. Modelos de aprendizagem supervisionada identificam a relação dos atributos com o rótulo da instância.

Assim, dado um conjunto de atributos que constituem um documento, é possível atribuir um rótulo (polaridade) através da classificação supervisionada. A ideia da técnica é representar o conjunto de dados como um vetor, que, em seguida, poderá ser usado pelo classificador para aprender e estimar rótulos. Diferentes tipos de métodos podem ser usados para treinar o classificador, mas o método mais comum e simples usado para classificação de textos é Naïve Bayes (PANG; LEE; VAITHYANATHAN, 2002; GO; BHAYANI; HUANG, 2009; GAUTAM; YADAV, 2014; TANG; KAY; HE, 2016).

O modelo é baseado no teorema de Bayes com a suposição de que os recursos são independentes. O classificador Naïve Bayes define a probabilidade do documento pertencente a uma determinada classe. As vantagens do classificador Bayes são simplicidade da implementação, processo de aprendizagem rápido e bons resultados (PANG; LEE; VAITHYANATHAN, 2002; GAUTAM; YADAV, 2014). No entanto, a suposição "ingênua" pode causar problemas dado que, nas características do mundo real, são há dependências.

O classificador de Entropia Máxima (*Maximum Entropy*) não assume a independência dos recursos. Assim, esse classificador teoricamente pode superar o Naïve Bayes. No entanto, o algoritmo de Entropia Máxima é mais difícil de implementar e o processo de aprendizagem é mais lento.

Outra abordagem para a classificação é baseada em regras. A ideia por trás do método é aplicar um conjunto de regras que foram geradas por especialistas com base na análise da área específica do domínio. Este método pode mostrar bons resultados ao usar uma ampla gama de regras. No entanto, a criação de tais regras é complexa e demorada.

A abordagem baseada em regras foi utilizada por Chikersal *et al.* (2015) juntamente com outra abordagem: *Support Vector Machine* (SVM).

O método assume uma divisão do espaço em subespaços que correspondem a classes particulares. Em termos de classificação binária, a ideia do estágio de treinamento é descobrir um hiperplano que separe melhor um conjunto de dados em duas classes com a máxima margem. A margem é a distância do hiperplano ao ponto de dados mais próximo do conjunto definido pelo hiperplano. Esses pontos de dados próximos do hiperplano são chamados de vetores de suporte. Estes últimos, são elementos críticos porque a remoção deles mudaria a posição do separador (MITCHELL, 1997).

Para concluir, SVM pode superar, em alguns casos, algoritmos como Naïve Bayes e *Maximum Entropy* (PANG; LEE; VAITHYANATHAN, 2002). No entanto, o SVM não é adequado para grandes conjuntos de dados devido ao tempo de treinamento poder tornar-se longo dependendo do número de exemplos e dimensionalidade dos dados.

Outra solução para classificação de textos é o uso de Redes Neurais Artificiais. A rede neural artificial segue os princípios da rede neural biológica. Supõe-se que a rede neural possa resolver problemas da mesma maneira que os humanos o fazem. As Redes Neurais consistem numa coleção de neurônios interligados, geralmente em várias camadas. A rede neural é capaz de aprender através do ajuste dos pesos dos neurônios (MITCHELL, 1997), e as mais comuns da literatura aplicadas à classificação de textos são as Redes Neurais Convolucionais e Redes Neurais Recorrentes (AGGARWAL; ZHAI, 2012).

A Árvore de Decisão é outra maneira de realizar a classificação, e consiste na decomposição hierárquica do espaço de dados. A estrutura da árvore contém dois tipos de nós: nó de folha (contém o valor do atributo de destino, ou seja, etiqueta positiva ou negativa na tarefa de classificação binária) e nó de decisão (contém uma condição em um dos atributos para divisão espacial). A partição do espaço de dados é feita de forma recursiva (AGGARWAL; ZHAI, 2012).

A segunda abordagem aplicada à Análise de Sentimentos é o **método baseado em léxico**, detalhado a seguir.

Este método utiliza-se de um léxico que consiste em termos com os respectivos índices de sentimento para cada termo. O termo pode ser associado a uma única palavra, frase ou idioma (TABOADA *et al.*, 2011; CHIAVETTA; LO BOSCO; PILATO, 2016). O sentimento é definido com base na presença ou ausência de termos no léxico. A abordagem baseada em léxico inclui abordagens baseadas em *corpus* e abordagem baseada em dicionário.

O raciocínio por trás da abordagem baseada em dicionário é usar bancos de dados lexicais com palavras de opinião para extrair o sentimento do documento (LIU, 2012). Um conjunto de palavras de sentimento (por exemplo, bom, ruim) com suas polaridades é coletado manualmente. O próximo passo é usar as palavras polares para enriquecer o conjunto pesquisando sinônimos e antônimos em um banco de dados lexical. Exemplos de tais bancos de dados são *WordNet*, *HowNet*, *SentiWordNet*, *SenticNet* e o MPQA (MILLER, 1995; MUSTO; SEMERARO; POLIGNANO, 2014).

O procedimento de pesquisa é iterativo. Em cada iteração, o algoritmo incorpora um conjunto de palavras atualizadas (conjunto expandido) e faz a busca novamente até não haver novas palavras a serem incluídas. Ao final, um conjunto de palavras de sentimento pode ser revisado com o objetivo de excluir eventuais erros.

Hu e Liu (2004) concentraram sua pesquisa na classificação de avaliações de clientes, ou seja, extraíram características de produtos que contêm sentimentos, depois

classificaram frases com base nessas características e, como resultado, o resumo das revisões do produto foi composto.

Por exemplo, se uma revisão fosse sobre uma câmera, os autores recuperaram recursos como a qualidade da imagem e o tamanho da câmera, e usando esses recursos, a classificação foi feita em avaliações de câmera positivas e negativas. Para atribuir uma marca positiva ou negativa para uma frase, primeiro, os pesquisadores recuperaram as palavras polares de cada revisão (HU; LIU, 2004).

Neste caso, adjetivos foram utilizados. A predição baseou-se na polaridade de um adjetivo que teve a mesma polaridade que seus sinônimos e oposto à polaridade de seus antônimos. As palavras polares foram utilizadas para pesquisar seus sinônimos e antônimos com orientação conhecida no dicionário *WordNet*. Portanto, a orientação das palavras polares que aparecem na revisão foi identificada. O método descrito pelos autores mostrou bons resultados, com precisão média 84% (HU; LIU, 2004).

Kim e Hovy (2004) investigaram o sentimento do texto e seu titular em relação a um determinado tópico. Os autores aplicaram vários classificadores. O primeiro classificador foi aplicado a cada palavra na frase para obter sua polaridade. O segundo classificador definiu a polaridade de toda a frase expressa pelo titular da opinião. Além disso, os autores apresentaram o uso de uma pequena lista inicial de palavras de semente (*seed*) de forma semelhante ao trabalho de Hu e Liu (2004) (usando adjetivo e verbos). Este último foi estendido procurando por sinónimos e antônimos oriundos do *WordNet* (MILLER, 1995).

Os autores mencionaram que alguns sinônimos e antônimos tinham uma orientação neutra ou mesmo oposta, o que os tornou inapropriados para uso. Além disso, os pesquisadores enfatizaram a necessidade de definir a força de positividade e negatividade das palavras que permitiriam eliminar palavras ambíguas (KIM; HOVY, 2004).

Kim e Hovy identificaram quatro regiões diferentes na frase que são próximas do suporte de opinião e podem conter sentimento. Para determinar a orientação das frases, os autores desenvolveram três modelos. O primeiro modelo baseou-se no pressuposto de que os negativos se cancelam um ao outro (KIM; HOVY, 2004).

O segundo e terceiro modelos foram a média harmônica e geométrica das forças de sentimento na região específica, respectivamente. Após a realização de experimentos concluiu-se que os melhores resultados obtidos foram usando o primeiro modelo e região que começa do titular da opinião até o final da frase (KIM; HOVY, 2004).

Park e Kim (2016) desenvolveram um método que usava três dicionários diferentes (normalmente apenas um é usado) para obter sinônimos e antônimos com base em palavras de

sementes. O léxico expandido foi utilizado para a classificação de *tweets*. Os autores afirmam que a técnica proposta possibilitou classificar *tweets* que o método tradicional baseado em dicionário não era capaz. No entanto, a abordagem sugerida tem várias desvantagens. O principal problema é a construção de uma coleção de sinônimos e antônimos que requerem muito tempo. Além disso, geralmente os dicionários contêm palavras formais, mas os *tweets* estão cheios de termos informais.

De um modo geral, a principal desvantagem da abordagem baseada em dicionário é a incapacidade de detectar palavras de sentimento com orientações de polaridade específicas de domínio e contexto (MEDHAT; HASSAN; KORASHY, 2014).

Bing Liu (2012) denota que a abordagem baseada em *corpus* pode ser aplicada em dois casos. O primeiro caso é a identificação de palavras de opinião e suas polaridades no *corpus* de domínio usando um determinado conjunto de palavras de opinião. O segundo caso de aplicação é visando construir um novo léxico dentro do domínio específico de outro léxico usando um *corpus* de domínio. Os resultados sugerem que, mesmo que as palavras de opinião dependam do domínio, pode acontecer que a mesma palavra tenha orientação oposta, dependendo do contexto.

Hazivassiloglou e McKeown (1997) debruçaram-se sobre a técnica baseada em *corpus*. Os autores propuseram um método que extraía a orientação semântica de adjetivos conjugados do corpus. A técnica é baseada no uso de letras textuais e palavras de opinião de semente (adjetivos). Regras linguísticas especiais são aplicadas aos corpos para descobrir palavras de opinião com polaridades correspondentes. Os autores assumem que os adjetivos têm a mesma polaridade se forem acompanhados pela conjunção "e". No entanto, a conjunção "mas" é usada para ligar adjetivos com polaridades opostas, ou seja, às vezes, essas regras não são aplicáveis (HATZIVASSILOGLOU; MCKEOWN, 1997).

Portanto, os autores também predizem as polaridades dos adjetivos unidos para verificar se as polaridades são iguais ou não. Para isso, o modelo de regressão log-linear é aplicado. Após o estágio de predição, obtém-se o gráfico que fornece links entre adjetivos. Em seguida, o agrupamento é realizado no gráfico para dividir adjetivos em subconjuntos positivos e negativos. Com este método Hazivassiloglou e McKeown conseguiram atingir 90% de precisão (HATZIVASSILOGLOU; MCKEOWN, 1997).

Conforme mencionado anteriormente, a mesma palavra de sentimento pode ter orientação semântica diferente dependendo do contexto. Ding *et al.* (2008) propuseram um método para encontrar a orientação do sentimento transmitida por revisores. Os autores

enfatizaram que alguns adjetivos, principalmente quantificadores, dependem do contexto e podem alterar suas polaridades.

Ding *et al.* (2008) usaram palavras, frases e idiomas como um léxico de opinião. A lista de adjetivos e advérbios é baseada em (HU; LIU, 2004) e ampliada pelos autores para incluir verbos e substantivos. Além disso, eles anotaram cerca de 1000 idiomas que contêm sentimentos claramente expressos. Depois concluído o léxico, eles definem a pontuação de polaridade para cada característica na frase de revisão (DING; LIU; YU, 2008).

Para obter a pontuação para a frase completa, os autores resumem todas as pontuações usando a função de pontuação proposta que dá melhores resultados do que a soma simples usada em (HU; LIU, 2004). Além disso, os autores aplicaram várias regras linguísticas para lidar com negações e orações que contêm a conjunção "mas" (DING; LIU; YU, 2008).

Além disso, o artigo apresenta uma abordagem holística para resolver o problema da polaridade de identificação das palavras de sentimento dependentes do contexto. Para este propósito, sugere-se três técnicas de consistência sobre conectividade: técnica de conjunção intra-oração, técnica de conjunção pseudo-intra-oração e técnica de conjunção entre orações. Para resumir, os autores relatam que a abordagem proposta é eficaz e dá melhores resultados do que os métodos propostos anteriormente (DING; LIU; YU, 2008).

O método baseado em *corpus* sozinho é menos eficaz que o método baseado em dicionário devido às limitações das palavras constantes no *corpus*. No entanto, o uso desta abordagem pode ajudar a construir um léxico específico de contexto e de domínio.

Em geral, o desempenho de métodos baseados em léxico em termos de complexidade e precisão do tempo depende fortemente do número de palavras no dicionário, ou seja, o desempenho diminui significativamente com o crescimento exponencial do tamanho do dicionário (THAKKAR; PATEL, 2015).

A abordagem estatística dedica-se a encontrar padrões de co-ocorrência ou palavras de opinião. Isso pode ser feito através da derivação de polaridades posteriores usando a co-ocorrência de adjetivos em um *corpus*, conforme proposto por Fahrni e Klenner (2008). É possível usar todo o conjunto de documentos indexados na web como o *corpus* para a construção do dicionário. Isso supera o problema da indisponibilidade de algumas palavras se o *corpus* usado não for suficientemente grande (TURNEY, 2002).

A polaridade de uma palavra pode ser identificada ao analisar-se a frequência de ocorrência da palavra em um *corpus* de textos anotados. Se a palavra ocorre mais frequentemente entre os textos positivos, sua polaridade é positiva, se ocorre mais

frequentemente entre os textos negativos, então é negativa. Se tiver frequências iguais, então é uma palavra neutra (READ; CARROLL, 2009).

As palavras de opinião semelhantes frequentemente aparecem juntas em um *corpus*. Se duas palavras aparecem juntas frequentemente dentro do mesmo contexto, é provável que tenham a mesma polaridade. Portanto, a polaridade de uma palavra desconhecida pode ser determinada calculando a frequência relativa de co-ocorrência com outra palavra (TURNEY, 2002).

A principal vantagem da abordagem estatística reside em sua simplicidade de implementação e o fato de as técnicas estatísticas não levarem em consideração o domínio dos dados. Sua maior desvantagem reside no fato de não ser capaz de identificar desambiguação de sentenças.

Este tópico forneceu explicações sobre algoritmos que podem ser aplicados à tarefa de classificação de textos. Além disso, algumas abordagens que foram utilizadas pelos pesquisadores para a classificação de texto são examinadas e mais especificamente, algumas abordagens baseadas em aprendizado de máquina e léxicos são discutidas.

Para os fins deste trabalho, a Análise de Sentimentos é aplicada com o objetivo de compreender a inclinação das opiniões dos clientes em relação a suas experiências em restaurantes, domínio abordado nesta pesquisa.

A abordagem escolhida foi a baseada em léxico e esta escolha deve-se a uma série de fatores, sobretudo a limitações relacionadas à quantidade de documentos a serem analisados.

A escolha parte do pressuposto que, devido ao fato de as empresas envolvidas neste trabalho serem Pequenas ou Médias, a quantidade de feedbacks talvez não seja grande o suficiente para a aplicação de técnicas que usam abordagens de aprendizagem de máquina, dado que tais técnicas demandam uma quantidade relativamente grande de dados para as atividades de treinamento e teste. Sendo assim, devido a esta limitação, métodos estatísticos ou baseados em dicionário tendem a ser mais recomendados (DING; LIU; YU, 2008; PANG; LEE, 2008; TABOADA *et al.*, 2011; CHIAVETTA; LO BOSCO; PILATO, 2016).

Além de compreender o sentimento geral das opiniões dos clientes em relação a suas experiências com os restaurantes abordados nesta pesquisa, faz-se necessário saber sobre o que os clientes mais falam quando elogiam ou criticam algum aspecto.

Para este fim, optou-se pela técnica de Modelagem de Tópicos como forma de extrair assuntos sobre os quais as pessoas mais falam, discutida a seguir como forma de extrair conhecimento a partir das opiniões de clientes extraídas das redes sociais das empresas que constituem o objeto desta pesquisa. A Modelagem de Tópicos consiste em uma técnica

probabilística não supervisionada, utilizada para descobrir, extrair e agrupar termos em coleções de estruturas temáticas, cujo funcionamento é explicado no tópico a seguir.

2.5 Modelagem de Tópicos

A escala atual de geração de conteúdo sob a forma de textos e sua ampla disponibilidade de acesso criaram demandas quanto à organização e classificação desse tipo de dado que não poderiam ser supridas por anotação humana. Em função desse limite, uma possível solução para tratar tal volume de dados baseia-se em técnicas de modelagem probabilística de tópicos, cujo principal objetivo é a descoberta de tópicos e a anotação de grandes coleções de documentos por classificação temática (BLEI, 2012).

Nos últimos anos os modelos de tópicos estatísticos emergiram como um método para descobrir tópicos de em grandes coleções de documentos de texto. A modelagem de tópicos é um método de aprendizado não supervisionado que assume que cada documento consiste em uma mistura de tópicos e cada tópico é uma distribuição de probabilidade sobre palavras. Um modelo de tópico é basicamente um modelo generativo de documentos que especifica um procedimento probabilístico pelo qual os documentos podem ser gerados. O resultado da modelagem de tópicos é um conjunto de *clusters* (grupos) de palavras. Cada *cluster* forma um tópico e é uma distribuição de probabilidade sobre palavras na coleção de documentos (LIU, 2012).

Tais técnicas empregam métodos estatísticos às palavras dos textos originais visando assim descobrir os temas presentes nestes, a relação destes temas entre si e ainda como evoluem ao longo do tempo. Os algoritmos de modelagem de tópicos não requerem nenhuma anotação ou classificação prévia dos documentos, sendo que os tópicos emergem da análise dos textos originais tal qual foram produzidos (LIU, 2012; BLEI, 2012).

As duas estratégias mais frequentes na literatura para a modelagem de tópicos são a *Probabilistic Latent Semantic Analysis* (pLSA) (HOFMANN, 1999), técnicas originada a partir da técnica *Latent Semantic Analysis* (LSA) e a *Latent Dirichlet Allocation* (LDA) (BLEI; NG; JORDAN, 2003).

A *Latent Semantic Analysis* (Análise Semântica Latente) é uma técnica patenteada por Deerwester *et al.* em 1988. Na LSA, uma redução de dimensionalidade projeta os documentos no chamado espaço semântico latente com o objetivo de encontrar relações

semânticas entre as entidades dadas (por exemplo, os documentos no *corpus*) (DEERWESTER *et al.*, 1990).

O principal objetivo da Análise Semântica Latente (LSA) é criar uma representação baseada em vetores para que os textos "façam conteúdo semântico". Por meio da representação vetorial a Análise Semântica Latente calcula a semelhança entre os textos para escolher as palavras relacionadas com o sucesso. A Análise Semântica Latente (LSA) usa Decomposição de Valor Singular (SVD) para reorganizar os dados (DEERWESTER *et al.*, 1990).

A Decomposição de Valor Singular é um método que usa uma matriz para reconfigurar e calcular todas as reduções do espaço vetorial. Além disso, as reduções no espaço vetorial serão computadas e organizadas por ordem de importância, da maior para a menos importante. Sendo assim, a suposição mais significativa será usada para encontrar o significado do texto, caso contrário, o menos importante será ignorado durante a suposição. Palavras que tenham uma alta taxa de similaridade ocorrem se essas palavras tiverem um vetor semelhante (DEERWESTER *et al.*, 1990).

Em 1999, Hofmann introduziu uma extensão para a Análise Semântica Latente (LSA) desenvolvida ao distanciar o modelo de abordagens algébricas e construí-lo em uma base estática, a saber, um modelo de classe latente (HOFMANN, 1999).

Para determinar a decomposição ideal na pLSA, a decomposição do valor singular é omitida e, em vez disso, minimiza a divergência de Kullback-Leibler entre a distribuição de probabilidade empírica derivada e a distribuição de probabilidade do modelo (HOFMANN, 2001).

A *Latent Dirichlet Allocation* (LDA) é um modelo generativo probabilístico e modelo de associação mista que foi introduzido por Blei *et al.* (2003), que descobre tópicos latentes em corpos de texto. Desde então, tornou-se um modelo muito usado no Processamento de Linguagem Natural.

Como modelo generativo e modelo de associação mista, a LDA postula que cada documento surge de uma mistura de documentos em vários graus (BLEI; NG; JORDAN, 2003; BLEI, 2012).

Um exemplo (aproximado) é o seguinte: o modelo determina que o documento #1729 é composto 37% sobre o *tópico 17*, 25% sobre o *tópico 11*, 14% sobre o *tópico 61*, etc. Enquanto o *tópico 17* está distribuído sobre as palavras: 34% Maçã, 21% banana, 4% laranja, 3,8% limão, etc. Pode-se derivar que o *tópico 17* é sobre ‘frutas’, mas, como a LDA não é supervisionada, esse rótulo não é usado, nem necessário (BLEI; NG; JORDAN, 2003).

Dada sua popularidade, facilidade de implementação e o fato de não demandar treinamento prévio, o modelo *Latent Dirichlet Allocation* (LDA) foi escolhido para ser aplicado a este trabalho. Portanto, faz-se necessário entender melhor seu funcionamento, o que se fará a seguir.

Para a *Latent Dirichlet Allocation* (LDA) os dados são tratados como oriundos de um processo generativo que contém variáveis ocultas (BLEI; NG; JORDAN, 2003). Esse processo define uma distribuição de probabilidade conjunta sobre as variáveis aleatórias observadas e as ocultas, que por sua vez é usada para calcular a distribuição condicional das variáveis ocultas dadas as variáveis observadas. Essa distribuição condicional também é chamada de distribuição posterior, considerando que problema computacional de inferir a estrutura de tópicos oculta a partir de um conjunto de documentos é o problema de calcular a distribuição posterior, ou seja, a distribuição condicional das variáveis ocultas dados os documentos (BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2009; BLEI, 2012).

O modelo assume que os tópicos são gerados antes dos documentos. Um tópico, por sua vez, é definido como uma distribuição de probabilidade sobre um vocabulário fixo. Como exemplo, um tópico sobre biologia será aquele que contém palavras relacionadas à biologia com maior probabilidade de ocorrência. Em contraposição, um tópico que se relacione com qualquer outro assunto distinto conterá palavras sobre biologia com probabilidade de ocorrência muito baixa ou zero (BLEI; NG; JORDAN, 2003). Todos os tópicos contêm distribuições com probabilidades sobre todo o vocabulário fixo, mas essas probabilidades só assumirão valores mais altos nos termos que caracterizam aquele tópico (BLEI, 2012).

O processo que gera os documentos em LDA é realizado em duas etapas. Para a geração de cada documento da coleção, tem-se:

1. Uma distribuição sobre tópicos é escolhida aleatoriamente.

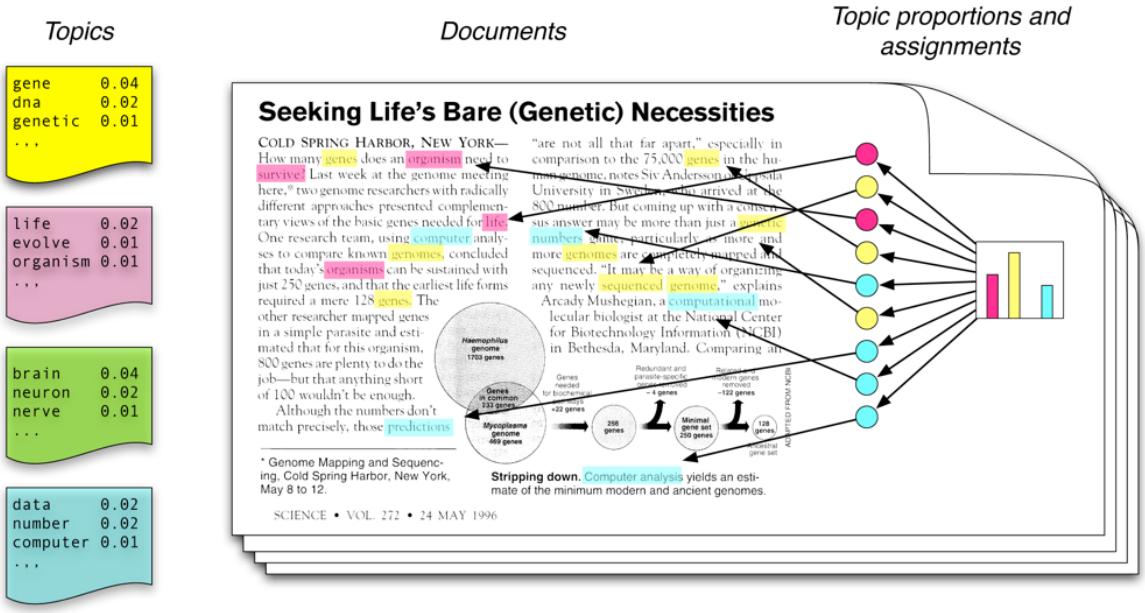
   Exemplo: Em um modelo com apenas 3 tópicos, uma distribuição possível sobre tópicos para um documento A pode exibir probabilidades 0.1, 0 e 0.9 de ocorrência dos tópicos x, y e z respectivamente.
2. Para cada palavra no documento:
   a) Um tópico é escolhido aleatoriamente a partir da distribuição obtida no passo 1.
   b) Uma palavra é escolhida aleatoriamente a partir do tópico (que por sua vez é uma distribuição de probabilidade sobre o vocabulário) obtido no passo 2.a.

Cada documento exibe tópicos em proporções distintas (passo 1), cada palavra em cada documento é obtida a partir de um dos tópicos (passo 2b), que por sua vez é escolhido a

partir da distribuição sobre tópicos de um documento em particular (passo 2a). Esse modelo reflete a intuição de que documentos exibem múltiplos tópicos, pressuposto por trás da formulação do modelo LDA (BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2009; BLEI, 2012).

No modelo LDA as variáveis observadas são as palavras nos documentos e as variáveis ocultas são a estrutura de tópicos, conforme demonstrado na Figura 5.

Figura 5 – Atribuição de tópicos a um documento em LDA



Fonte: Adaptado de Blei (2012).

O modelo da LDA descobre a estrutura latente de tópicos revertendo o processo gerativo formalizado. A partir das informações observadas – ou seja, os padrões de co-ocorrência na distribuição de palavras do conjunto de documentos do *corpus* – o modelo infere a estrutura de tópicos e sua distribuição a partir do modelo gerativo (BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2009; BLEI, 2012).

O resultado da aplicação desta técnica permite a exploração da estrutura latente de tópicos presente num *corpus*. Neste trabalho, a LDA é aplicada para extrair conhecimento a partir de opiniões de clientes, retornando uma quantidade de tópicos e os termos mais prováveis em cada um dos tópicos.

A aplicação prática da técnica será detalhada no Capítulo 3 - Procedimentos Metodológicos, e seus resultados serão discutidos no Capítulo 4 - Apresentação e Análise dos Resultados.

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia é a "aplicação de procedimentos e técnicas que devem ser observados para construção do conhecimento, com o propósito de comprovar sua validade e utilidade nos diversos âmbitos da sociedade" (PRODANOV; FREITAS, 2013, p. 14). Com base neste conceito, este capítulo é dedicado a descrever o rigor do caminho metodológico aplicado nesta pesquisa, detalhando as escolhas do estudo, visando assim garantir a base necessária à sua confiabilidade e replicabilidade.

3.1 Classificação da pesquisa

O objetivo da metodologia é o "aperfeiçoamento dos procedimentos e critérios utilizados na pesquisa", e o método, por sua vez, é "o caminho para se chegar a determinado fim ou objetivo" (MARTINS; THEÓPHILO, 2017, p. 35).

Segundo Marconi e Lakatos (2003, p. 155) a pesquisa científica "é um procedimento formal, com método de pensamento reflexivo, que requer um tratamento científico e se constitui no caminho para conhecer a realidade ou para descobrir verdades parciais".

Assim sendo, esta pesquisa apresenta-se como pesquisa científica que busca avançar o conhecimento existente em relação à aplicação da mineração de textos para descoberta de conhecimento do cliente referente a experiências em restaurantes oriundas de redes sociais aplicável à realidade de pequenas e médias empresas.

No que diz respeito à sua natureza, esta pesquisa caracteriza-se como uma pesquisa aplicada e descritiva. De acordo com Gil (2008a, p. 27), a pesquisa aplicada "tem como característica fundamental o interesse na aplicação, utilização e consequências práticas dos conhecimentos". Para Prodanov e Freitas (2013, p. 51), a pesquisa aplicada "objetiva gerar conhecimentos para aplicação prática dirigidos à solução de problemas específicos". Em complemento, é descritiva pois versa sobre aspectos como a descrição, registro, análise e interpretação de um problema (GIL, 2008b).

Para a realização desta pesquisa foi empregado o método indutivo, à medida que a "aproximação dos fenômenos caminha geralmente para planos cada vez mais abrangentes, indo das constatações mais particulares às leis e teorias" (LAKATOS; MARCONI, 2003, p. 106). Assim sendo, a pesquisa que se utiliza do método indutivo parte da observação de fatos ou fenômenos cujas causas se quer conhecer (GIL, 2008b).

Em relação à abordagem, esta pesquisa é definida como pesquisa qualitativa, dado que há uma relação dinâmica entre o mundo real e o sujeito, isto é, um vínculo indissociável entre o mundo objetivo e a subjetividade do sujeito que não pode ser traduzido em números (SILVA; MENEZES, 2005; PRODANOV; FREITAS, 2013). Não obstante, esta pesquisa emprega métodos quantitativos para o desenvolvimento de algumas de suas etapas, seja quantificando, sumarizando ou usando técnicas de estatística descritiva para explorar relações entre os dados, na busca por evidenciar seu significado num contexto específico.

Pesquisas qualitativas não apresentam aversão à quantificação de variáveis, mas enfatizam a captação das perspectivas e interpretações dos indivíduos estudados. Nas pesquisas qualitativas o foco está no entendimento de um determinado fenômeno, produto de interpretação e significados a ele atribuídos pelo pesquisador, e não na frequência com que este fenômeno ocorre, dado que "a interpretação dos fenômenos e a atribuição de significados são básicas no processo de pesquisa qualitativa" (SILVA; MENEZES, 2005, p. 20).

Segundo Strauss e Corbin (1998), métodos qualitativos podem ser usados para explorar áreas nas quais o conhecimento existente é pequeno, ou aplicados à áreas nas quais o conhecimento é expressivo, como forma de proporcionar novos pontos de vista.

Na visão de Marconi e Lakatos (2003), a escolha adequada do instrumental metodológico está diretamente relacionada ao problema a ser estudado, o que implica que tal escolha depende dos fatores relacionados ao estudo proposto. Assim, métodos e técnicas devem adequar-se ao problema, às questões de pesquisa e ao objeto a ser abordado, podendo, por vezes constituir-se de uma combinação de dois ou mais deles, empregados concomitantemente.

Face ao exposto, o método experimental foi escolhido para a condução desta pesquisa. Para Gil (2008b), a pesquisa experimental consiste em determinar um objeto de estudo, selecionar as variáveis capazes de influenciá-lo e definir as formas de controle e de observação dos efeitos que a variável produz no objeto. Desta forma, o escopo da pesquisa experimental volta-se à extração do conhecimento do universo, no caso a base de documentos que se constitui das opiniões dos clientes. Para tanto, foram aplicadas técnicas de mineração de textos e apresentação dos resultados por meio de análises qualitativas e descritivas. Esta

pesquisa visa ainda apresentar uma nova abordagem em relação aos procedimentos para a mineração de textos em opiniões de clientes dispostas em redes sociais.

Martins e Theóphilo (2017, p. 35) argumentam que "a expressão 'método científico', tal qual é empregada contemporaneamente, pode induzir a crer que consiste em regras exaustivas e infalíveis", quando na verdade, não existem tais receitas para conduzir-se uma investigação científica. Os autores afirmam também que "o que há são estratégias de investigação científica com técnicas gerais e particulares, e métodos especiais para diversas tecnologias e ciências" (MARTINS; THEÓPHILO, 2017, p. 35).

Assim sendo, apresenta-se a seguir o *Framework* para Mineração de Opiniões que empregado nesta pesquisa, bem como a abordagem e as ferramentas empregadas para tal finalidade.

3.2 *Framework* para mineração de opiniões

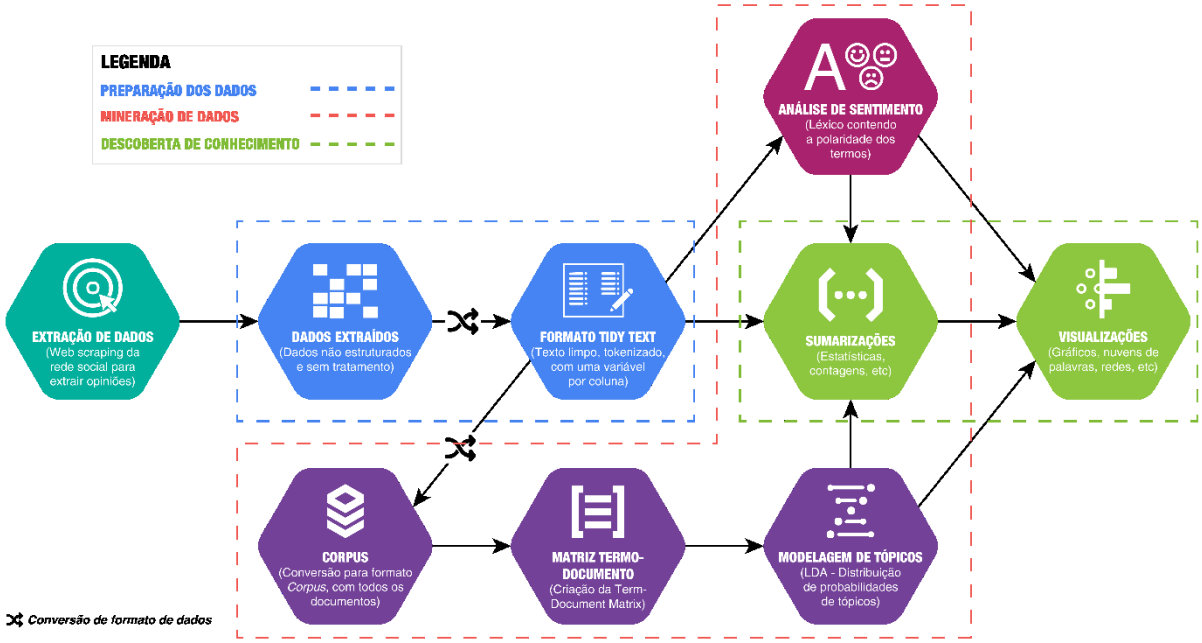
Neste tópico são analisadas e discutidas as etapas de desenvolvimento do *framework* para mineração de opiniões para descoberta de conhecimento do cliente referente a suas experiências em restaurantes oriundas de redes sociais, aplicável à realidade de pequenas e médias empresas.

Por tratar-se de uma área em desenvolvimento, muitas pesquisas sugerem metodologias e abordagens diferentes para a mineração de textos. Assim, não há um consenso sobre quais etapas devem ou não ser implementadas, havendo quem defenda que a decisão sobre quais etapas realizar deve ser avaliada no contexto de cada aplicação específica (ARANHA, 2007; FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2017).

Convém observar que o objetivo da mineração e o domínio no qual ela ocorre tem muita relevância sobre a escolha da metodologia em si e não somente das etapas a serem desempenhadas, como ainda das técnicas que envolvem cada uma etapa (FELDMAN; SANGER, 2007).

Considerando-se o objetivo que norteia esta pesquisa, apresenta-se na Figura 6 a proposta de *framework* para mineração de opiniões que busca descobrir conhecimento a partir de opiniões de clientes de restaurantes publicadas na rede social TripAdvisor Brasil.

Figura 6 – *Framework* de Mineração de Opiniões de clientes



Fonte: Elaborado pelo autor.

O *framework* para mineração de opiniões apresentada na Figura 6 baseia-se nas propostas formuladas por Aranha (2007) e Feldman e Sanger (2007), ampliando-as com a aplicação de técnicas adicionais voltadas à Análise de Sentimento (DING; LIU; YU, 2008; LIU, 2012; PANG; LEE, 2008) e à Modelagem de Tópicos (BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2009; BLEI, 2012), ambas previamente apresentados no capítulo de Fundamentação Teórica.

O emprego de Análise de Sentimentos e Modelagem de Tópicos justifica-se nesta pesquisa em função do objetivo geral delineado, no caso, a extração de conhecimento das opiniões dos clientes por meio da aplicação de técnicas de Mineração de Textos. A aplicação de ambas no contexto desta pesquisa é melhor explicada e discutida no decorrer deste capítulo.

O *framework* de mineração de opiniões adotada para esta pesquisa apoia-se numa abordagem de mineração de dados conhecida como *Tidy Data* (dados arrumados [tradução nossa]) (WICKHAM, 2014) que, por sua vez, apoia-se num ecossistema de ferramentas baseado nos mesmos princípios. A abordagem *Tidy Data* e as respectivas ferramentas a ela associadas, bem como sua aplicação no modelo proposto para esta pesquisa são apresentadas a seguir.

3.3 Abordagem *tidy data* para mineração de textos

Dados se perfazem na matéria-prima para que processos de mineração ocorram, e por natureza, os dados podem ser estruturados e não estruturados (SILVA; PERES; BOSCARIOLI, 2017). Independentemente de sua natureza, todo dado requer algum tipo de preparo antes de sua análise, o que por vezes é descrito como um trabalho que requer muito esforço por parte dos cientistas de dados, principalmente em se tratando de dados não estruturados, como é o caso de textos, por exemplo (FELDMAN; SANGER, 2007).

Dasu e Johnson (2003) afirmam que cerca de 80% do tempo da análise de dados são gastos no processo de limpeza e preparação dos dados brutos. Em relação ao tratamento de dados, um relatório publicado em 2016 pela CrowdFlower (empresa especializada em ciência de dados), revela que 76% dos cientistas de dados veem a preparação de dados como a parte menos agradável de seu trabalho. O mesmo relatório afirma que cientistas de dados gastam 60% do tempo em limpeza e organização de dados. A coleta de conjuntos de dados é a segunda atividade mais dispendiosa, com 19% do tempo, o que significa que os cientistas de dados gastam cerca de 80% de seu tempo na preparação e gerenciamento de dados para análise (CROWDFLOWER, 2016).

Isto significa que as etapas que antecedem a mineração de dados e a descoberta de conhecimento em si acabam por consumir muito mais tempo do que as análises em si, o que pode impactar diretamente em uma série de fatores, tais como custos e eficiência de processos. Assim, a preparação de dados não é apenas um primeiro passo, mas deve ser repetida muitas vezes ao longo da análise à medida que novos problemas aparecem ou novos dados são coletados (WICKHAM, 2014).

Uma possível solução para lidar com o problema relatado é a adoção de uma abordagem em relação aos dados que permita com que os processos de limpeza e organização tomem menos tempo, sobretudo quando se trata de dados não estruturados, como é o caso de textos, cuja complexidade tende a ser maior (SILGE; ROBINSON, 2017; SILVA; PERES; BOSCARIOLI, 2017).

Wickham (2014) propõe um tipo de estruturação de conjuntos de dados para facilitar a análise denominada *tidy data*, que nada mais é do que a aplicação de uma série de princípios que visa garantir a organização de valores num conjunto de dados. O padrão *tidy data* foi projetado para facilitar a exploração e análise inicial dos dados, propiciando ainda a

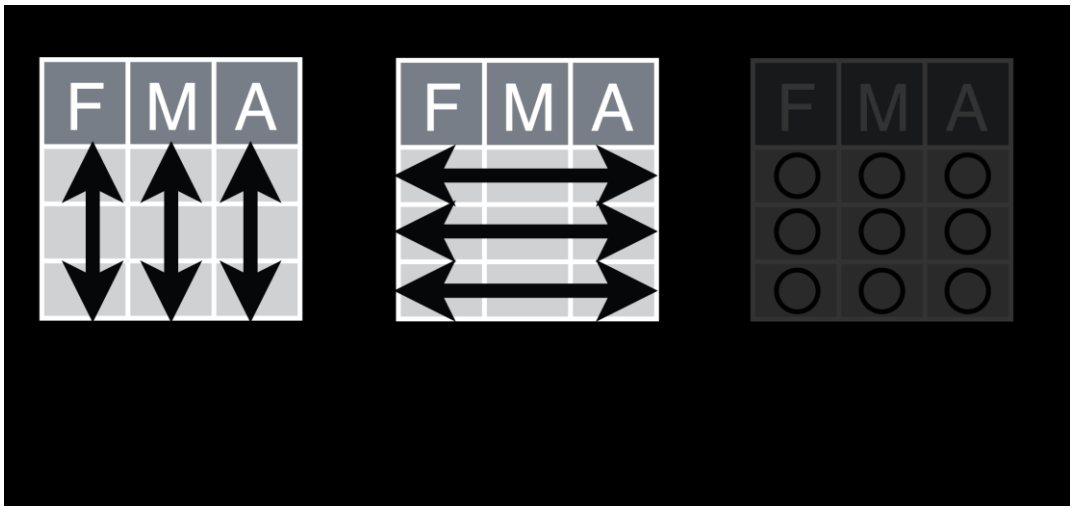
simplificação do desenvolvimento de ferramentas de análise de dados que funcionem bem trabalhando juntas. O autor ressalta alguns dos principais motivos da complexidade de lidar com dados:

1. Os nomes de colunas representam valores de dados em vez de nomes de variáveis;
2. Uma única coluna contém dados em diversas variáveis, ao invés de uma única variável;
3. As variáveis estão contidas em linhas e colunas, ao invés de apenas colunas;
4. Uma única tabela contém mais de uma unidade de observação;
5. Os dados sobre uma unidade de observação estão espalhados por vários conjuntos de dados.

Pyle (1999) afirma que o trabalho de exploração dos dados não começa com os dados em si, mas em sua preparação. A preparação de dados é, portanto, um tema de pesquisa crucial, não obstante muitos trabalhos no campo da mineração de dados pressuporem a existência de dados de qualidade antes da mineração, ou seja, dados onde a entrada para os algoritmos de mineração é supostamente distribuída, não contendo valores ausentes ou incorretos (ZHANG; ZHANG; YANG, 2003).

O uso dos princípios de *tidy data* é uma forma robusta de tornar a manipulação de dados mais fácil e efetiva. Conforme descrito por Wickham (2014), os princípios do *tidy data* estão intimamente ligados aos princípios de bancos de dados relacionais, bem como aos princípios de álgebra relacional de Codd (1990). Sobre este último aspecto, mais especificamente à terceira forma normal proposta por Codd, uma vez que o *tidy data* possui uma estrutura específica, que consiste em basicamente três regras: cada variável forma uma coluna; cada observação forma uma linha e cada tipo de unidade de observação forma uma tabela. A Figura 7 ilustra como os princípios do *tidy data* devem ser aplicados aos conjuntos de dados.

Figura 7 – Estrutura específica para dados arrumados



Fonte: Baseado em (WICKHAM, 2014).

Aplicam-se estes princípios à mineração de textos, objeto desta pesquisa. Silge e Robinson (2017) definem o formato *tidy text* como sendo uma tabela com um *token* por linha. O *token* é uma unidade significativa de texto (normalmente uma palavra) que se está interessado em analisar. Assim, *tokenization* (ou tokenização) é o processo de dividir um texto em *tokens* (FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2017). Essa estrutura de um *token* por linha contrasta com as formas pelas quais o texto é normalmente armazenado nas análises de texto: como *strings* ou em uma matriz de documento-termo.

Para realizar a mineração de texto arrumado, o *token* que está armazenado em cada linha é geralmente uma única palavra, mas também pode ser um *n-gram*, uma frase ou um parágrafo. O código em linguagem R do Quadro 4 mostra a criação de um vetor de caracteres aplicado a um trecho de poema de Manoel de Barros, que mais adiante será tratado usando o princípio de dados arrumados.

Quadro 4 – Código para criação de um vetor de caracteres

```
# Criando um vetor de caracteres
poema <- c(
  "A maior riqueza do homem",
  "é sua incompletude.",
  "Nesse ponto sou abastado.",
  "Palavras que me aceitam como sou",
  "– eu não aceito."
)

# Exibindo o vetor
poema

## [1] "A maior riqueza do homem"     "é sua incompletude."
## [3] "Nesse ponto sou abastado."    "Palavras que me aceitam como sou"
## [5] "– eu não aceito."
```

Fonte: Elaborado pelo autor.

Este é um vetor de caracteres típico, que apresenta um formato comum ao início de qualquer processo de mineração de textos. Para transformá-lo em um conjunto de dados de texto arrumado (*tidy data*), é necessário convertê-lo num formato de dados tabular, normalmente um *data.frame*. Esse formato de dados é muito comum na linguagem R, assemelhando-se ao formato de dados dispostos numa planilha eletrônica (SILVA; PERES; BOSCARIOLI, 2017; WICKHAM; GROLEMUND, 2016). O código apresentado no Quadro 5 mostra o processo de conversão do vetor gerado no passo anterior em uma estrutura conhecida como *data.frame*.

Quadro 5 – Conversão do vetor em uma estrutura do tipo *data.frame*

```
# Carregando o pacote "dplyr"
library(dplyr)

# Criando um data.frame a partir do vetor
poema_df <- data_frame(line = 1:5, poema = poema)

# Visualizando o data.frame
poema_df

## # A tibble: 5 x 2
##    line                              poema
##   <int>                              <chr>
## 1     1         A maior riqueza do homem
## 2     2              é sua incompletude.
## 3     3        Nesse ponto sou abastado.
## 4     4 Palavras que me aceitam como sou
## 5     5                 – eu não aceito.
```

Fonte: Elaborado pelo autor.

Como resultado da criação do *data.frame*, uma estrutura chamada *tibble* foi criada. Uma estrutura *tibble* é uma classe moderna de *data.frames* na linguagem R (WICKHAM; GROLEMUND, 2016) e possui uma série de vantagens sobre os *data.frames* convencionais, o que a torna ideal para ser usada em mineração de dados arrumados (*tidy data*). Um detalhe que deve ser observado é a criação de um índice para cada linha do poema, no exemplo exposto. Isto torna-se necessário dada a realidade dos dados com os quais se desenvolve este trabalho, no caso, um documento por linha. A criação deste índice será muito útil no próximo: tokenização, explicado a seguir.

O próximo passo no processo é transformar cada linha do *data-frame* no formato final de análise que consiste em um *token* por linha. Para isso, emprega-se um comando do pacote *tidytext* chamado *unnest_tokens* (SILGE; ROBINSON, 2017), conforme indica o Quadro 6.

Quadro 6 – Processo de tokenização, resultando em um *token* por linha

```
# Carregando o pacote "dplyr"
library(dplyr)

# Criando um data.frame a partir do vetor
poema_df <- data_frame(line = 1:5, poema = poema)

# Visualizando o data.frame
poema_df

## # A tibble: 5 x 2
##    line                              poema
##   <int>                              <chr>
## 1     1         A maior riqueza do homem
## 2     2              é sua incompletude.
## 3     3        Nesse ponto sou abastado.
## 4     4 Palavras que me aceitam como sou
## 5     5                 – eu não aceito.
```

Fonte: Elaborado pelo autor

Ao final da operação de criação de *tokens*, o resultado é um *data.frame* do tipo *tibble* no formato de um *token* por linha e, neste caso, dividida em duas colunas: uma contendo a linha de onde o *token* foi retirado (que funciona como um índice, como explicado anteriormente) e a segunda coluna contendo o *token* em si.

A partir daqui, todas as operações realizadas com este *data.frame* são, tipicamente, baseadas em operações entre colunas e tabelas, de forma muito semelhante ao que acontece em bancos de dados relacionais (WICKHAM, 2014). A Figura 8 mostra como seria uma operação realizada entre colunas de dois *data.frames* diferentes.

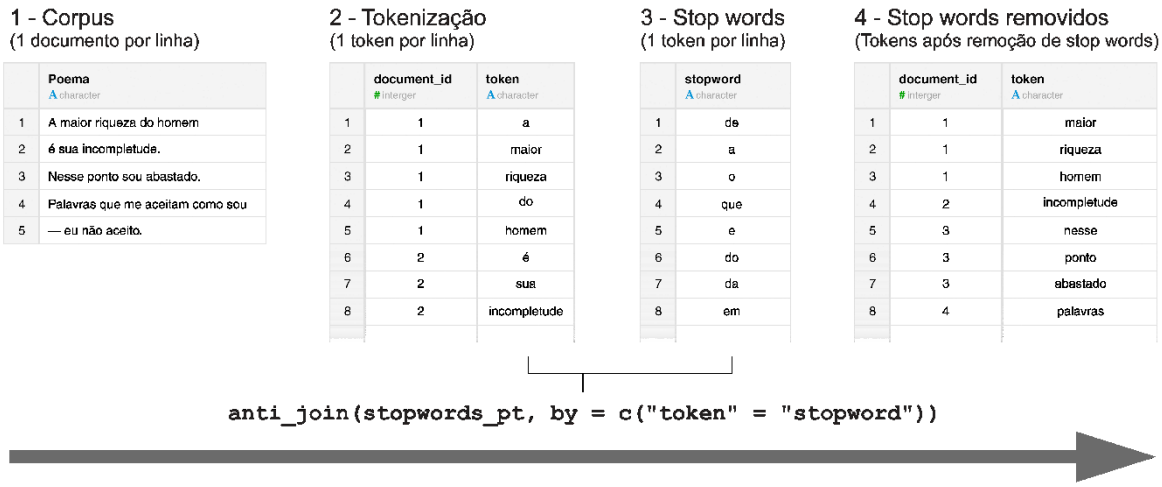
Figura 8 – Uma operação *"inner_join"* entre duas colunas de diferentes *data.frames*

Fonte: Elaborado pelo autor.

Pode-se aplicar o mesmo princípio às operações comuns a todo processo de mineração de textos como, por exemplo, a remoção de *stop words*, que são termos que não contribuem para a análise por não terem em si nenhum valor semântico (FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2017).

Seguindo este mesmo princípio, e considerando que a lista de *stop words* pode ser tratada da mesma forma que a análise de sentimentos foi, a operação de remoção de *stop words* pode ser realizada por meio de um comando *anti_join*, que retorna todas as linhas de 'A' onde não há valores correspondentes em 'B', mantendo apenas as colunas de 'A', conforme mostra a Figura 9.

Figura 9 – Processo de remoção de *stop words* em dados arrumados



Fonte: Adaptado de Silge e Robinson (2017).

O mesmo princípio pode ser aplicado a todas as operações de mineração de textos como, por exemplo, a normalização de caracteres maiúsculas de minúsculas, operações de

limpeza e remoção de números e caracteres especiais, como *emojis* e até mesmo operações mais sofisticadas, como a análise de sentimentos.

Neste caso, uma operação como a análise de sentimentos baseada em léxico, uma das abordagens tratadas nesta pesquisa, aconteceria seguindo o esquema representado na Figura 10, que usa uma operação do tipo *inner_join* entre duas colunas de diferentes *data.frames* (SILGE; ROBINSON, 2017). Considerando-se que o léxico usado está no mesmo formato que o conteúdo de texto, ou seja, um *token* por linha, a operação seria tal qual demonstrada na Figura 10.

Figura 10 – Processo de análise de sentimentos em dados arrumados

Fonte: Adaptado de Silge e Robinson (2017).

Diversas operações podem ser realizadas com dados arrumados, como será apresentado no decorrer desta dissertação. A grande vantagem em realizar a mineração de textos seguindo este princípio está na dinâmica e simplicidade das operações, o que proporciona análises rápidas e agiliza operações como contagens, filtragens e manipulações que envolvem o cruzamento de diferentes colunas de dados (SILGE; ROBINSON, 2017; WICKHAM; GROLEMUND, 2016; WICKHAM, 2014).

Para realizar todas as operações contidas na metodologia de Mineração de Textos proposta nesta pesquisa, uma série de ferramentas foi empregada nas diferentes etapas desenvolvidas. No próximo tópico descreve-se em detalhes cada ferramenta usada nas diferentes etapas desta pesquisa.

3.4 Ferramentas empregadas

Com a finalidade de implementar o método apresentado no tópico anterior foram empregadas ferramentas que executam: (1) a extração das opiniões de usuários de restaurantes publicadas na rede social TripAdvisor Brasil, (2) o pré-processamento dos textos com o intuito de formatar e organizar de maneira adequada para a mineração dos dados e (3) a mineração dos dados e apresentação de visualizações gráficas. A seguir são descritas as ferramentas empregadas em cada uma das etapas do *framework* aplicado nesta pesquisa.

Nos últimos anos a área de Ciência de Dados tem adquirido muito interesse, tanto por parte da Academia, quanto pelo mercado, o que se reflete no aumento exponencial de ferramentas, tecnologias e técnicas para a análise de dados (HOFMANN; CHISHOLM, 2013). Para o desenvolvimento desta pesquisa havia diversas possibilidades, incluindo o emprego de soluções prontas (softwares ou interfaces gráficas) e diversas linguagens de programação. Após a análise comparativa das diversas alternativas aventadas, definiu-se os seguintes critérios para a escolha da tecnologia a ser empregada neste experimento:

1. **Livre acesso** – As ferramentas empregadas devem ser de uso livre, dado que se trata de um trabalho acadêmico, cujo foco não comercial torna proibitivo o emprego de ferramentas que sejam pagas em função do consequente ônus para reproduções futuras do método proposto;
2. **Reprodutibilidade** – Este critério tem a ver com o critério anterior, dado que, ao optar-se por uma ferramenta paga, a reprodução dos resultados obtidos neste experimento seria dificultada. Além disso, diversos aspectos da reprodutibilidade do método escolhido foram testados, de modo a viabilizar os futuros experimentos.
3. **Flexibilidade** – A ferramenta escolhida deve ser flexível a ponto de ser utilizada sem necessidade de customizações complexas, estando isenta das limitações típicas de soluções prontas.
4. **Facilidade de uso** – Considerando o tempo exíguo para ao desenvolvimento desta pesquisa, a facilidade de uso (interface amigável) ou implementação da solução a ser adotada torna-se um critério importante na seleção das ferramentas para o desenvolvimento do experimento.

A análise das soluções disponíveis para mineração de dados reduziu-se a duas alternativas: a linguagem Python e a linguagem R. Ambas satisfazem todos os pré-requisitos definidos anteriormente. Elas possuem diversas semelhanças entre si, como o fato de trabalharem com o conceito de *packages* (pacotes), possuírem bases de usuários vastas e repositórios de pacotes igualmente repletos de ferramentas validadas por diversos trabalhos acadêmicos e aplicações de mercado.

Neste experimento optou-se pelo uso da linguagem R em função do aumento expressivo da quantidade de usuários que passaram a adotar a linguagem R no dia a dia. Em artigo publicado no site *Stack Overflow*, o cientista de dados David Robinson analisa o crescimento da linguagem R em relação a outras linguagens de programação por meio de dados extraídos do próprio site *Stack Overflow*, evidenciando que nos últimos anos, nenhuma linguagem de programação teve crescimento mais expressivo que a linguagem R (ROBINSON, 2017). Assim, a escolha por essa linguagem também se deu em função das perspectivas de aplicação futura desta ferramenta.

Outro motivo da escolha da linguagem R para o desenvolvimento dos experimentos desta pesquisa tem a ver com a própria natureza desta linguagem, criada inicialmente para lidar com dados estatísticos (WICKHAM; GROLEMUND, 2016). Esta pesquisa emprega uma série de manipulações estatísticas com diferentes níveis de complexidade. Assim, o uso da linguagem R facilita e simplifica de forma significativa várias etapas deste processo, como será visto no decorrer da dissertação.

A principal ferramenta empregada nesta pesquisa foi a linguagem de programação R '*Short Summer'* (versão 3.4.2 de 28/set/2017), uma linguagem de programação livre e que pode ser baixada na Internet a partir do site <http://cran.r-project.org>. Neste experimento, a linguagem R foi operada por meio do RStudio (RSTUDIO TEAM, 2015) versão 1.0.153, que é um ambiente de desenvolvimento integrado (IDE) para a linguagem R. Esta IDE inclui um console, um editor de destaque de sintaxe que suporta a execução direta de código, bem como ferramentas para plotar gráficos e ainda ferramentas para depuração e gerenciamento de espaço de trabalho.

Parte da popularidade da linguagem R junto à comunidade acadêmica deve-se ao fato de que vários pesquisadores desenvolvem e disponibilizam gratuitamente bibliotecas que executam dezenas de tipos de análises. Tais bibliotecas também são conhecidas como pacotes (*packages*), que podem ser baixados e carregados localmente, conforme as necessidades do pesquisador. Neste experimento empregaram-se diversos pacotes que cumprem tarefas específicas, cujas funções são descritas a seguir.

Para a extração dos dados das páginas de restaurantes selecionados no site TripAdvisor Brasil, o pacote ***rvest*** (versão 0.3.2) foi usado (WICKHAM, 2016). O tópico seguinte (Coleta e pré-processamento dos dados) descreve em detalhes como foram realizados a extração e o tratamento dos dados.

Para esta segunda etapa foram utilizados vários pacotes que desempenharam funções específicas, tendo sido empregados em paralelo. A primeira biblioteca utilizada foi a ***tidyverse*** (WICKHAM, 2017) na versão 1.1.1, que contém os pacotes expostos no Quadro 7[1], com destaque para os itens que foram usados nesta pesquisa, exibidos em negrito.

Quadro 7 – Pacotes carregados pela biblioteca *"tidyverse"*

Importação	Arrumar (Tidy)	Transformar	Programar	Modelar	Visualizar
readr ***readxl*** haven httr ***rvest*** xml2	***tibble*** ***tidyr***	***dplyr*** forcats hms ***lubridate*** ***stringr***	purrr ***magrittr***	broom modelr	***ggplot2***

Fonte: Elaborado pelo autor.

O *tidyverse* é um sistema coerente de pacotes para manipulação, exploração e visualização de dados que compartilha de uma filosofia de *design* comum. Estes pacotes foram desenvolvidos em sua maioria por Hadley Wickham e mais recentemente vem sendo expandidos pela comunidade da Ciência de Dados. Fundamentalmente, a filosofia por trás do *tidyverse* tem a ver com as conexões entre as ferramentas, o que torna o fluxo de trabalho mais fluído. O pacote *tidyverse* carrega vários subpacotes, conforme exposto no Quadro 7, porém, nas fases de normalização, mineração e apresentação de dados, apenas os pacotes *readr, readxl rvest, tibble, tidyr, lubridate, stringr, magrittr* e *ggplot2* foram aplicados nesta pesquisa e são descritos a seguir, juntamente com as demais bibliotecas empregadas no decorrer deste trabalho.

As bibliotecas ***readr*** e ***readxl*** fornecem maneiras rápidas e amigáveis de ler dados retangulares (como arquivos csv, tsv, xls e xlsx) e foram projetadas para analisar de forma flexível muitos tipos de arquivos diferentes.

Quanto às biliotecas ***dplyr*** e ***tidyr***, a primeira é um dos pacotes mais úteis para realização da manipulação de dados, permitindo filtrar, organizar, criar subconjuntos, modificar e agregar dados em *data frames*, tendo sido aqui usada juntamente com o *tidyr*, que é uma

[1] Os elementos destacados em negrito foram utilizados nesta pesquisa.

biblioteca de funções projetada especificamente para a arrumação de dados (*tidy data*, já explicado anteriormente).

Já quanto as bibliotecas ***reshape2*** e ***stringr*** usadas neste experimento: o *reshape2* serviu para a conversão dos dados, o que facilita a transformação de dados entre formatos amplos e longos. O pacote *stringr* foi usado para manipular *strings* de forma eficiente.

O pacote ***tibble*** nada mais é do que um *data.frame*, porém com um método de impressão mais adequado. Praticamente todas as bibliotecas que fazem parte do *tidyverse* produzem *tibbles* que, por sinal, tem muito a ver com a ideia de tratar os dados segundo a abordagem *tidy data*, explicada em detalhes anteriormente.

O pacote ***magrittr*** oferece o operador '%>%', também conhecido como *pipe*, que serve para concatenar comandos e executá-los em sequência. Basicamente, o operador '%>%' usa o resultado do seu lado esquerdo como primeiro argumento da função do lado direito, o que torna a execução de comandos muito mais simples e de fácil leitura. Neste experimento, o operador '%>%' foi amplamente usado, possibilitando assim executar cadeias de comando de forma rápida.

Relativamente aos pacotes ***ggplot2***, ***wordcloud, igraph*** e ***ggraph***: o pacote *ggplot2* foi usada para gerar parte dos gráficos de visualização dos dados, pois trata-se de um pacote voltada à criação de gráficos estatísticos (WICKHAM, 2009). Outra parte das visualizações foi gerada em formato de nuvens de palavras, o que foi feito com o auxílio do pacote *wordcloud* (FELLOWS, 2014). Além disso, os pacotes *igraph* (CSARDI; NEPUSZ, 2006) e *ggraph* (PEDERSEN, 2017) foram usados para gerar visualizações de redes de palavras.

O pacote ***tidytext*** desempenhou papel fundamental em parte do experimento realizado. Este pacote contém uma série de funções para lidar com mineração de textos usando os princípios de dados arrumados (*tidy data*) (SILGE; ROBINSON, 2017).

O pacote ***ptstem*** contém três algoritmos de *stemming* para a língua portuguesa: algoritmo de *Porter*, *Hunspell* e *RSLP* (FALBEL, 2017). O *stemming*, apesar de aplicado neste trabalho, não foi levado até o final por dois motivos: não havia necessidade de redução de termos ao radical, dado o número pequeno de documentos analisados e o fato de que a análise de sentimentos é feita comparando a coluna de *tokens* com uma coluna de léxicos, ambos completos, o que inviabiliza o uso de *stemming* nas análises. Ainda assim, o *stemming* foi feito e guardado em uma variável separada para possíveis análises.

O pacote ***topicmodels*** inclui interfaces para dois algoritmos para a modelagem de tópicos (GRÜN; HORNIK, 2011). Neste experimento fez-se uso da técnica de modelagem de

tópicos de David M. Blei (BLEI, 2012), pelo fato de ser a implementação mais comum na literatura (BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2009; BLEI, 2012).

Além dos pacotes e bibliotecas apresentados, fez-se uso de vários comandos provenientes do próprio ambiente de desenvolvimento do R, como por exemplo o *install.packages()*, usado no processo de instalação de pacotes e o comando *library()*, usado para carregar pacotes antes de sua utilização, e doravante referenciados como comandos ***r-core***, por fazerem parte do núcleo da instalação do R. O Quadro 8 mostra um resumo de todas as ferramentas empregadas neste trabalho.

Quadro 8 – Visão geral das ferramentas empregadas nesta pesquisa

Linguagem	Ambiente	Bibliotecas	Software Adicional
R	RStudio	tidyverse tidytext rvest wordcloud igraph ggraph ptstem topicmodels r-core	Microsoft Excel

Fonte: Elaborado pelo autor.

Uma vez explicadas as ferramentas aplicadas no desenvolvimento deste trabalho, serão abordados a seguir os detalhes referentes à coleta e tratamento dos dados, desde sua limpeza, até a sua apresentação.

3.5 Coleta e pré-processamento dos dados

Os dados utilizados neste experimento foram extraídos diretamente da rede social TripAdvisor Brasil (TRIPADVISOR, 2017), escolhida para a realização desta pesquisa por conter opiniões de usuários sobre suas experiências diversos tipos de empreendimentos, dentre os quais se destacam hotéis, restaurantes e empresas prestadoras de serviços.

A fase de extração de dados, para efeitos de representação, é apresentada a seguir na Figura 11. O processo de extração de dados da rede social, denominado *web scraping* (MUNZERT, 2015), é detalhado mais adiante.

Figura 11 – Esquema do processo de extração dos dados do site TripAdvisor

Fonte: Elaborado pelo autor.

O TripAdvisor é uma rede social na qual os usuários descrevem e avaliam suas experiências nos mais diversos locais, tanto comerciais (hotéis e restaurantes), como turísticos (atrações, museus e parques, dentre outros). Muitas empresas brasileiras estão presentes nesta rede social, embora algumas delas sequer tenham conhecimento sobre sua presença, enquanto outras a ignoram. Isto deve-se porque, assim como ocorre com outras redes sociais tais como Facebook ou Google My Business, as páginas das empresas podem ser criadas tanto de forma intencional, ou seja, pela própria empresa, como à revelia desta, o que ocorre quando as páginas são criadas pelas pessoas que frequentam estas empresas ou locais.

Atualmente, o TripAdvisor Brasil conta com mais de 3.600 restaurantes cadastrados no país (TRIPADVISOR, 2017). Há de se considerar, no entanto, que nem todos aproveitam essa presença para interagir com seus utilizadores. O TripAdvisor permite ao dono do empreendimento registrar-se como representante de um determinado local, dando-lhe a possibilidade de responder publicamente aos comentários, alterar informações sobre a organização, adicionar imagens e até ter acesso às estatísticas da página daquele estabelecimento.

Os indicadores do TripAdvisor justificam a escolha deste site como fonte de opiniões de consumidores sobre suas experiências com estabelecimento, notadamente quanto a restaurantes e hotéis. Com mais de 535 milhões de opiniões que cobrem a maior seleção mundial de listagens de viagens em todo o mundo (mais de 7 milhões de acomodações, companhias aéreas, atrações e restaurantes), o TripAdvisor fornece aos usuários a sabedoria das multidões para ajudá-los a decidir para onde viajar, onde ficar, o que fazer e onde se alimentar.

Os sites da marca do TripAdvisor estão disponíveis em 49 mercados e compõem a maior comunidade de viagens do mundo, com 415 milhões de visitantes mensais únicos em média (TRIPADVISOR, 2017).

Dadas as características e relevância do TripAdvisor, a escolha desta rede social como principal fonte de dados para esta pesquisa justifica-se pelo fato de ser uma rede dedicada a receber e prover indicações de pessoas que frequentam lugares como restaurantes, o que representa uma excelente oportunidade de extrair e avaliar opiniões voluntárias de usuários sobre produtos, serviços e experiências.

Redes como o Facebook também já foram avaliadas em pesquisas acadêmicas, porém, dado o fato de que redes como o Facebook não são dedicadas exclusivamente a opiniões de consumidores sobre produtos e serviços, a seleção destas opiniões torna-se mais difícil, muito embora não se negue sua maior popularidade e abrangência. A seguir, os passos que envolveram a coleta e pré-processamento dos dados extraídos do TripAdvisor Brasil são detalhados.

Os dados considerados nesta pesquisa consistem em opiniões de usuários sobre suas experiências com vários tipos de empreendimentos. Esta pesquisa foca especificamente as opiniões dadas em restaurantes, cuja seleção foi guiada pelos critérios definidos pelos seguintes parâmetros:

1. Restaurantes com preços populares;
2. Que pertençam à categoria ‘Hamburguerias’;
3. Ranqueados no site de classificações TripAdvisor Brasil (www.tripadvisor.com.br) com no mínimo 500 avaliações postadas;
4. Enquadrar-se na classificação de Pequena ou Média Empresa[2];
5. Não pertencer a uma rede de restaurantes ou franquia;
6. Restaurantes estabelecidos no município de São Paulo (SP).

A escolha do tipo de restaurante considerado nesta pesquisa justifica-se pelo público-alvo majoritariamente jovem e, portanto, mais propenso ao uso de redes sociais (CASEY, 2017). Após avaliação preliminar, dezenas de restaurantes se enquadraram no perfil acima indicado (aproximadamente 90 estabelecimentos). Como critério complementar, o refinamento da seleção foi feito considerando-se a quantidade de avaliações (*posts*) disponível.

[2] Ver Quadro 1 – Classificação de empresas segundo o SEBRAE, apresentado no Capítulo 1.

Quatro restaurantes que cumprem os critérios foram selecionados considerando a quantidade de avaliações disponíveis.

Visando uma melhor acurácia e representatividade dos resultados da pesquisa, decidiu-se por não restringir o intervalo de publicações de opiniões. Esta decisão deve-se, sobretudo, ao fato de a rede social TripAdvisor não ser ainda tão popular no Brasil quanto é nos Estados Unidos e Europa, o que implica numa massa de dados exponencialmente menor de avaliações feitas em língua portuguesa.

É importante ressaltar que o site TripAdvisor Brasil aceita publicações feitas em qualquer idioma, porém, para os fins desta pesquisa, optou-se pela extração e uso apenas de opiniões em língua portuguesa. Por fim, explicita-se que um dos critérios para a seleção da rede TripAdvisor Brasil tem a ver com sua disponibilidade pública e, portanto, acessível para o pesquisador por meio de técnicas de extração simples, ou seja, sem necessidade de acesso a bases privadas. Consta da literatura uma série de trabalhos acadêmicos que se utilizaram de dados extraídos diretamente do TripAdvisor (KEATES, 2007; MIGUÉNS; BAGGIO; COSTA, 2008; O'CONNOR, 2008; SMYTH; WU; GREENE, 2010; FERNANDES, 2015; ALJALIDI; ALSHEDOKHI; SABA, 2016; GATICA-PEREZ; RUIZ-CORREA; SANTANI, 2016).

Dos quatro restaurantes selecionados a partir dos critérios já descritos, o que possuía menor quantidade de avaliações foi tratado como EXPERIMENTO PRELIMINAR. Seus dados foram utilizados para criar e refinar o modelo de Mineração de Textos que foi posteriormente aplicado aos demais restaurantes. Os três restaurantes restantes, doravante denominados EMPRESA 1, EMPRESA 2 e EMPRESA 3, foram analisados em vista do modelo desenvolvido a partir do resultado dos dados do EXPERIMENTO PRELIMINAR, cuja quantidade de documentos extraída é apresentada no Quadro 9. Mais detalhes sobre cada uma das empresas abordadas nesta pesquisa serão apresentados no capítulo "Apresentação e Análise dos Resultados".

Quadro 9 – Empresas selecionadas para a pesquisa

Empresa	Quantidade de opiniões extraídas (documentos)	Intervalo de publicação das opiniões
EXPERIMENTO PRELIMINAR	555	06/2015 a 10/2017
EMPRESA 1	816	01/2012 a 10/2017
EMPRESA2	975	12/2009 a 10/2017
EMPRESA3	1292	12/2012 a 10/2017

Fonte: Elaborado pelo autor.

Para realizar a extração dos dados do site TripAdvisor Brasil empregou-se um pacote chamado ***rvest*** (WICKHAM, 2016), conforme já apresentado anteriormente. Este pacote foi criado por Hadley Wickham em 2016 para facilitar o processo conhecido como *web scraping*, que consiste na extração de dados de websites (MUNZERT, 2015). A extração dos dados ocorreu entre os dias 16 e 20 de outubro de 2017, tendo sido realizada em partes divididas em grupos de dez páginas, visando evitar problemas operacionais como bloqueio de IP.

As opiniões extraídas do site TripAdvisor Brasil seguem um formato não muito consistente. Alguns usuários deixam informações pessoais como seu Estado (unidade da federação) de origem e outras informações baseadas em suas opiniões como, por exemplo, uma escala de pontos para alguns critérios de avaliação do restaurante, tais como custo-benefício, atendimento e comida, enquanto outros deixam somente suas opiniões em forma de texto, mas nenhum dado ou opinião adicional.

Visando a uniformidade da análise, optou-se por extrair somente os dados que fossem comuns a todos os usuários do site, mesmo que uma parte dos dados originais disponíveis não fosse utilizada nas análises. Assim, de todos os campos existentes, os seguintes foram escolhidos para a extração:

- Nome do usuário (campo *NomeReviewer*);
- Nota dada pelo usuário ao restaurante (campo *NotaReviewer*);
- Data da criação da opinião (campo *DataReview*);
- Título da opinião (campo *TituloReview*);
- Texto da opinião (campo *TextoReview*).

À título de demonstração, a Figura 12 exibe como a opinião sobre um restaurante é disposta no site do TripAdvisor Brasil, destacando-se os campos escolhidos para o processo de extração, conforme citado acima.

Figura 12 – Uma opinião no TripAdvisor Brasil, destacando os campos extraídos

Fonte: Elaborado pelo autor.

O *script* usado para a extração das opiniões de usuários de restaurantes do site TripAdvisor Brasil está disponível no Apêndice A. O script de extração aplicado gera como resultado final um arquivo .CSV (*comma-separated values*), que nada mais é que uma implementação particular de arquivos de texto separados por um delimitador, que usa vírgulas e quebras de linha para separar valores.

A escolha por este formato de arquivo se deu pelo fato de que diferentes etapas desta pesquisa foram desenvolvidas em sistemas operacionais diferentes, o que torna a adoção de formatos universais imperativa para a replicação do experimento. Este arquivo CSV pode ser lido por qualquer programa de planilha eletrônica, o que possibilita facilmente sua edição. A Figura 13 mostra o aspecto do conteúdo de um arquivo .CSV, quando exibido em um software de leitura de texto simples.

Figura 13 – Conteúdo de um arquivo do tipo CSV com valores separados por vírgulas

```
DadosTripAdvisor__NomeReviwer;DadosTripAdvisor__TituloReview;DadosTripAdvisor__DataReview;DadosTripAdvisor__NotaReview;DadosTripAdvisor__TextoR
eview
Ronaldo1958;Almoço;8 de Outubro de 2017;ui_bubble_rating bubble_40;Excepcional, ambiente agradável hambúrguer de primeira qualidade... Garçons
atenciosos... Super recomendado!!!
santosfernanda55;Almoço;5 de Outubro de 2017;ui_bubble_rating bubble_20;Almocei hambúrguer com batata frita ..... e só posso dizer que me
decepcionei com a batata super gordurosa e murcha. O hambúrguer estava okay, embora tenham exagerado na quantidade de alface. Talvez eu tenha
criado muitas expectativas já que me indicaram como um excelente local. Achei os garçons pouco receptivos e precisava ficar procurando-os pois
nunca havia nenhum próximo  ♀
Ariane C;Hambúrguer saboroso;3 de Outubro de 2017;ui_bubble_rating bubble_40;O lugar é bem moderno e aconchegante! Além dos lanches há opção de
comida (modesto self service). O preço é um pouquinho salgado, mas compensa!
Vinidandrea;Bom lanche em SP;30 de Setembro de 2017;ui_bubble_rating bubble_40;O local passou por uma mudança visual muito boa, ficou mais
moderno e agradável . O lanche mudou um pouquinho o sabor mas continua muito bom.
Matheus B;Adorei o lugar;27 de Setembro de 2017;ui_bubble_rating bubble_50;Fui uma vez apenas ao chico, mas voltarei com certeza !!! Tem
bastante fila aos finas de semana então se prepare para esperar um pouco, estacionamento é gratuito !!! A lugar é grande e com muitas mesas, os
garçons são ágeis e atenciosos, o hambúrguer estava uma delícia no ponto certo que pedimos (ponto menos), a batata sequinha e crocante, a
maionese é aquela verde tipica de hamburgueria, muito boa !!! O preço vale o custo x benefício, você come bem e sai satisfeito !!!!
SusaneLopes;Mais opções, delicioso e, agora mais bonito!;20 de Setembro de 2017;ui_bubble_rating bubble_40;Com a reforma ficou um pouco mais
escondido, porem, muito mais moderno e agradavel. O cardapio está mais completo e, a qualidade sempre excelente! Como um chesse filet mignon
salada deliciosooo. Preservaram a qualidade! Parabens!
Eriel_Mineli;Vale a visita!;11 de Setembro de 2017;ui_bubble_rating bubble_40;Hambúrguer é um assunto muito particular! Sempre passo em frente
ao Chico Hambúrguer de Moema e sempre ensaio uma parada. Hoje, voltando de um Compromisso resolvi parar. O lugar é sensacional, porém o
atendimento deixa um pouco a desejar Como não conhecia a casa, esperava ao menos que o garçom me explicasse como funcionava o cardápio. Para
minha surpresa ele me entregou o cardápio, virou as costas e saiu. Chamei-o novamente e perguntei como funcionava e ele na menor da boa vontade
me explicou. Talvez o domingo tenha sido puxado pra eles e por ser segunda estavam cansados. Quanto ao Hambúrguer, estava correto, o ponto da
carne tal qual como pedi é muito saboroso. Não é barato, mas vale a pena conhecer, afinal, é como eu disse, Hambúrguer é muito pessoal.
Hambúrguer de Picanha + água + 10% = R$ 42,00.
Tânia C;Cheese-Tomate?!;4 de Setembro de 2017;ui_bubble_rating bubble_10;Sim, hoje fui surpreendida por um cheese-tomate, pois meu lanche tinha
pedaços enormes de tomate e quase nada de carne, que estava super engordurada. Não consegui terminar o lanche. Meu marido também. Que pena! A
batata estava mole e gordurosa, ou seja, ruim também! O único item que se salvou foi a Coca-Cola. Não vale o que cobram! Melhor ir na
concorrência.
```

Fonte: Elaborado pelo autor.

O resultado final da extração das opiniões do site TripAdvisor Brasil consistiu em uma estrutura que contém os campos definidos para extração em colunas e uma opinião por linha, formato condizente com os princípios dos dados arrumados (*tidy data*) proposto por Wickham (2014), conforme mostra a Figura 14 mais adiante.

Após a fase de extração dos dados do site TripAdvisor Brasil, iniciou-se a fase de **pré-processamento**, que compreende a limpeza e organização dos dados para que seja possível realizar a mineração de textos. Esta fase foi dividida em duas etapas: a) limpeza e estruturação dos dados e 2) normalização dos dados.

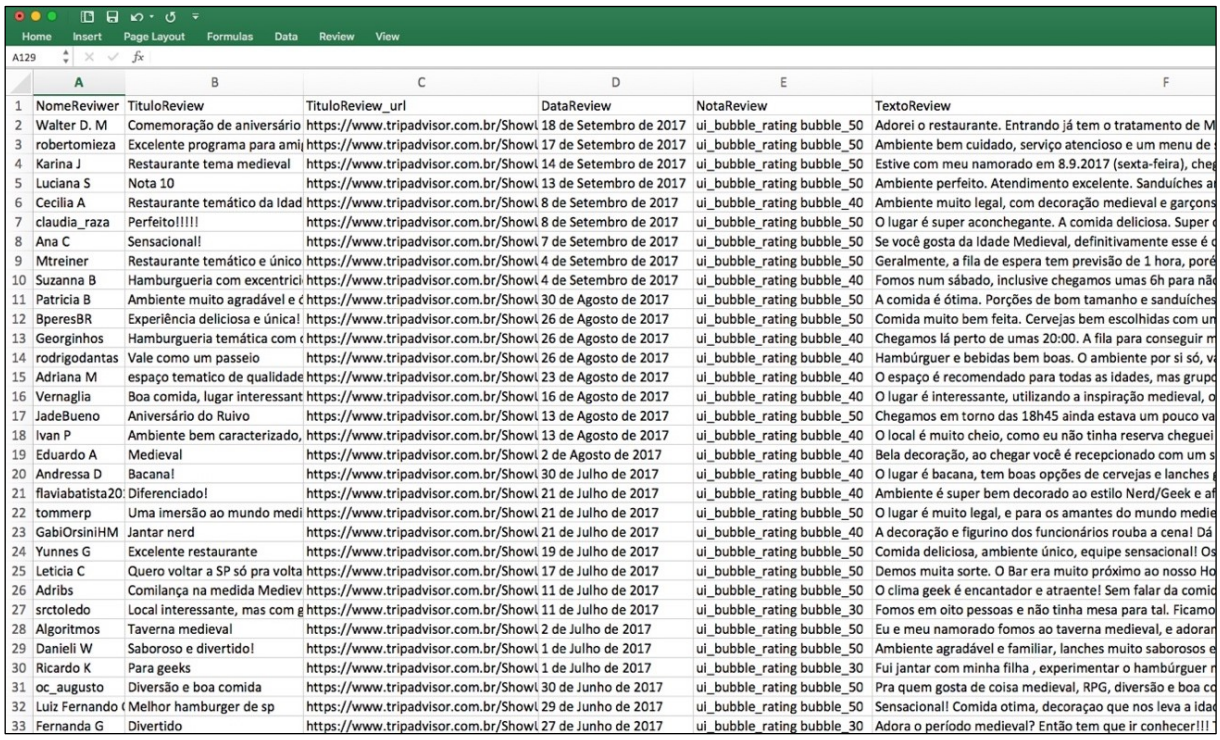
O intuito em separar o pré-processamento em duas etapas justifica-se pelo fato de que os dados, tal qual foram capturados do site do TripAdvisor Brasil, trazem diversas informações que não servem à análise pretendida neste experimento, tais como as colunas contendo URLs e nomes de usuários autores das opiniões. Isto tem a ver com uma característica típica ao tratamento de dados não estruturados que, na maioria das vezes, necessitam de diversos processos de limpeza e organização visando prepará-los para a fase de mineração (SILVA; PERES; BOSCARIOLI, 2017).

A primeira etapa do pré-processamento, aqui denominada ‘limpeza e estruturação dos dados’, ocorreu por meio de software externo. Já a segunda etapa executada no ambiente da linguagem R. Para realizar a limpeza e estruturação dos dados foi empregada a ferramenta Microsoft Excel (versão 17.0).

Embora esta etapa pudesse ser desempenhada diretamente no ambiente R, optou-se por usar uma ferramenta externa por questão de agilidade. Outro fator determinante para a

escolha do Microsoft Excel nesta etapa é o fato de que a linguagem R é capaz de ler praticamente qualquer tipo de arquivo, o que significa a não obrigação de se realizar todo a fase de pré-processamento no ambiente da linguagem R. A Figura 14 mostra como os dados importados sem tratamento são arranjados antes de iniciar-se a fase de limpeza e estruturação dos dados.

Figura 14 – Dados sem tratamento antes de iniciar-se a fase de limpeza e estruturação



	A	B	C	D	E	F
1	NomeReviwer	TituloReview	TituloReview_url	DataReview	NotaReview	TextoReview
2	Walter D. M	Comemoração de aniversário	https://www.tripadvisor.com.br/Showl	18 de Setembro de 2017	ui_bubble_rating bubble_50	Adorei o restaurante. Entrando já tem o tratamento de M
3	robertomieza	Excelente programa para ami	https://www.tripadvisor.com.br/Showl	17 de Setembro de 2017	ui_bubble_rating bubble_50	Ambiente bem cuidado, serviço atencioso e um menu de
4	Karina J	Restaurante tema medieval	https://www.tripadvisor.com.br/Showl	14 de Setembro de 2017	ui_bubble_rating bubble_50	Estive com meu namorado em 8.9.2017 (sexta-feira), cheg
5	Luciana S	Nota 10	https://www.tripadvisor.com.br/Showl	13 de Setembro de 2017	ui_bubble_rating bubble_50	Ambiente perfeito. Atendimento excelente. Sanduíches a
6	Cecilia A	Restaurante temático da Idad	https://www.tripadvisor.com.br/Showl	8 de Setembro de 2017	ui_bubble_rating bubble_40	Ambiente muito legal, com decoração medieval e garçons
7	claudia_raza	Perfeito!!!!!	https://www.tripadvisor.com.br/Showl	8 de Setembro de 2017	ui_bubble_rating bubble_50	O lugar é super aconchegante. A comida deliciosa. Super
8	Ana C	Sensacional!	https://www.tripadvisor.com.br/Showl	7 de Setembro de 2017	ui_bubble_rating bubble_50	Se você gosta da Idade Medieval, definitivamente esse é
9	Mtreiner	Restaurante temático e único	https://www.tripadvisor.com.br/Showl	4 de Setembro de 2017	ui_bubble_rating bubble_50	Geralmente, a fila de espera tem previsão de 1 hora, poré
10	Suzanna B	Hamburgueria com excentrici	https://www.tripadvisor.com.br/Showl	4 de Setembro de 2017	ui_bubble_rating bubble_40	Fomos num sábado, inclusive chegamos umas 6h para não
11	Patricia B	Ambiente muito agradável e c	https://www.tripadvisor.com.br/Showl	30 de Agosto de 2017	ui_bubble_rating bubble_50	A comida é ótima. Porções de bom tamanho e sanduíches
12	BperesBR	Experiência deliciosa e única!	https://www.tripadvisor.com.br/Showl	26 de Agosto de 2017	ui_bubble_rating bubble_50	Comida muito bem feita. Cervejas bem escolhidas com un
13	Georginhos	Hamburgueria temática com c	https://www.tripadvisor.com.br/Showl	26 de Agosto de 2017	ui_bubble_rating bubble_40	Chegamos lá perto de umas 20:00. A fila para conseguir m
14	rodrigodantas	Vale como um passeio	https://www.tripadvisor.com.br/Showl	26 de Agosto de 2017	ui_bubble_rating bubble_40	Hambúrguer e bebidas bem boas. O ambiente por si só, va
15	Adriana M	espaço tematico de qualidade	https://www.tripadvisor.com.br/Showl	23 de Agosto de 2017	ui_bubble_rating bubble_40	O espaço é recomendado para todas as idades, mas grupo
16	Vernaglia	Boa comida, lugar interessant	https://www.tripadvisor.com.br/Showl	16 de Agosto de 2017	ui_bubble_rating bubble_40	O lugar é interessante, utilizando a inspiração medieval, o
17	JadeBueno	Aniversário do Ruivo	https://www.tripadvisor.com.br/Showl	13 de Agosto de 2017	ui_bubble_rating bubble_50	Chegamos em torno das 18h45 ainda estava um pouco va
18	Ivan P	Ambiente bem caracterizado,	https://www.tripadvisor.com.br/Showl	13 de Agosto de 2017	ui_bubble_rating bubble_40	O local é muito cheio, como eu não tinha reserva cheguei
19	Eduardo A	Medieval	https://www.tripadvisor.com.br/Showl	2 de Agosto de 2017	ui_bubble_rating bubble_40	Bela decoração, ao chegar você é recepcionado com um s
20	Andressa D	Bacana!	https://www.tripadvisor.com.br/Showl	30 de Julho de 2017	ui_bubble_rating bubble_40	O lugar é bacana, tem boas opções de cervejas e lanches
21	flaviabatista20	Diferenciado!	https://www.tripadvisor.com.br/Showl	21 de Julho de 2017	ui_bubble_rating bubble_40	Ambiente é super bem decorado ao estilo Nerd/Geek e af
22	tommerp	Uma imersão ao mundo medi	https://www.tripadvisor.com.br/Showl	21 de Julho de 2017	ui_bubble_rating bubble_50	O lugar é muito legal, e para os amantes do mundo medie
23	GabiOrsiniHM	Jantar nerd	https://www.tripadvisor.com.br/Showl	21 de Julho de 2017	ui_bubble_rating bubble_40	A decoração e figurino dos funcionários rouba a cena! Dá
24	Yunnes G	Excelente restaurante	https://www.tripadvisor.com.br/Showl	19 de Julho de 2017	ui_bubble_rating bubble_50	Comida deliciosa, ambiente único, equipe sensacional! Os
25	Leticia C	Quero voltar a SP só pra volta	https://www.tripadvisor.com.br/Showl	17 de Julho de 2017	ui_bubble_rating bubble_50	Demos muita sorte. O Bar era muito próximo ao nosso Ho
26	Adribs	Comilança na medida Mediev	https://www.tripadvisor.com.br/Showl	11 de Julho de 2017	ui_bubble_rating bubble_50	O clima geek é encantador e atraente! Sem falar da comic
27	srctoledo	Local interessante, mas com g	https://www.tripadvisor.com.br/Showl	11 de Julho de 2017	ui_bubble_rating bubble_30	Fomos em oito pessoas e não tinha mesa para tal. Ficamo
28	Algoritmos	Taverna medieval	https://www.tripadvisor.com.br/Showl	2 de Julho de 2017	ui_bubble_rating bubble_50	Eu e meu namorado fomos ao taverna medieval, e adoran
29	Danieli W	Saboroso e divertido!	https://www.tripadvisor.com.br/Showl	1 de Julho de 2017	ui_bubble_rating bubble_50	Ambiente agradável e familiar, lanches muito saborosos e
30	Ricardo K	Para geeks	https://www.tripadvisor.com.br/Showl	1 de Julho de 2017	ui_bubble_rating bubble_30	Fui jantar com minha filha , experimentar o hambúrguer r
31	oc_augusto	Diversão e boa comida	https://www.tripadvisor.com.br/Showl	30 de Junho de 2017	ui_bubble_rating bubble_50	Pra quem gosta de coisa medieval, RPG, diversão e boa co
32	Luiz Fernando (	Melhor hamburger de sp	https://www.tripadvisor.com.br/Showl	29 de Junho de 2017	ui_bubble_rating bubble_50	Sensacional! Comida otima, decoraçao que nos leva a ida
33	Fernanda G	Divertido	https://www.tripadvisor.com.br/Showl	27 de Junho de 2017	ui_bubble_rating bubble_30	Adora o período medieval? Então tem que ir conhecer!!!

Fonte: Elaborado pelo autor.

O arquivo de dados no formato .CSV foi importado para o Microsoft Excel e os seguintes processos foram realizados visando limpar e estruturar os dados: 1) Eliminação de colunas sem utilidade para a análise; 2) Conversão do formato de data e 3) Conversão do formato da nota que cada usuário atribuiu ao negócio. Assim, as colunas eliminadas foram *NomeReviewer*, que continha os nomes dos usuários autores das opiniões e a *TituloReview_url*, que continha o link da página de onde as opiniões foram extraídas (no caso, sempre a mesma para todas as opiniões coletadas).

Além de eliminar colunas desnecessárias à análise, também realizou-se a conversão dos dados de duas outras colunas. A primeira foi a coluna *DataReview*, cujo formato original explícito impossibilita seu processamento durante a mineração de dados. Os formatos de data explícitos foram convertidos em formato de data segundo o padrão internacional ISO 8601

(ISO, 2004). Seguindo este princípio, o dado original ‘26 de Janeiro de 2017’ ficou normatizado como ‘26/01/2017’, respeitando-se o padrão ‘dd/mm/aaaa’.

Por último, a coluna *NotaReview* teve seu conteúdo convertido num formato numérico propício ao processo de mineração. Originalmente, a coluna *NotaReview* possui um conteúdo que representa uma *string* concatenada junto com a nota atribuída pelo usuário, resultando em algo como *“ui_bubble_rating bubble_40”*, que significa ‘nota 4’. Isso acontece por causa da maneira como esta informação é disponibilizada no TripAdvisor Brasil, que no site é exibida em formato de escala visual que vai de 1 a 5 ‘bolinhas’, cujo preenchimento é feito considerando a pontuação proveniente das notas dadas pelos clientes e exibidas por meio de uma combinação de *tags* HTML e estilos CSS.

Considerando-se a limitação de que a informação da nota não é disponibilizada de forma explícita, foi necessário extraí-la no formato em que estava disponível no site, que no caso trata-se de uma classe CSS atribuída a uma *tag* HTML *span*. A Figura 15 mostra como a escala funciona, bem como o código HTML e CSS responsável por sua exibição.

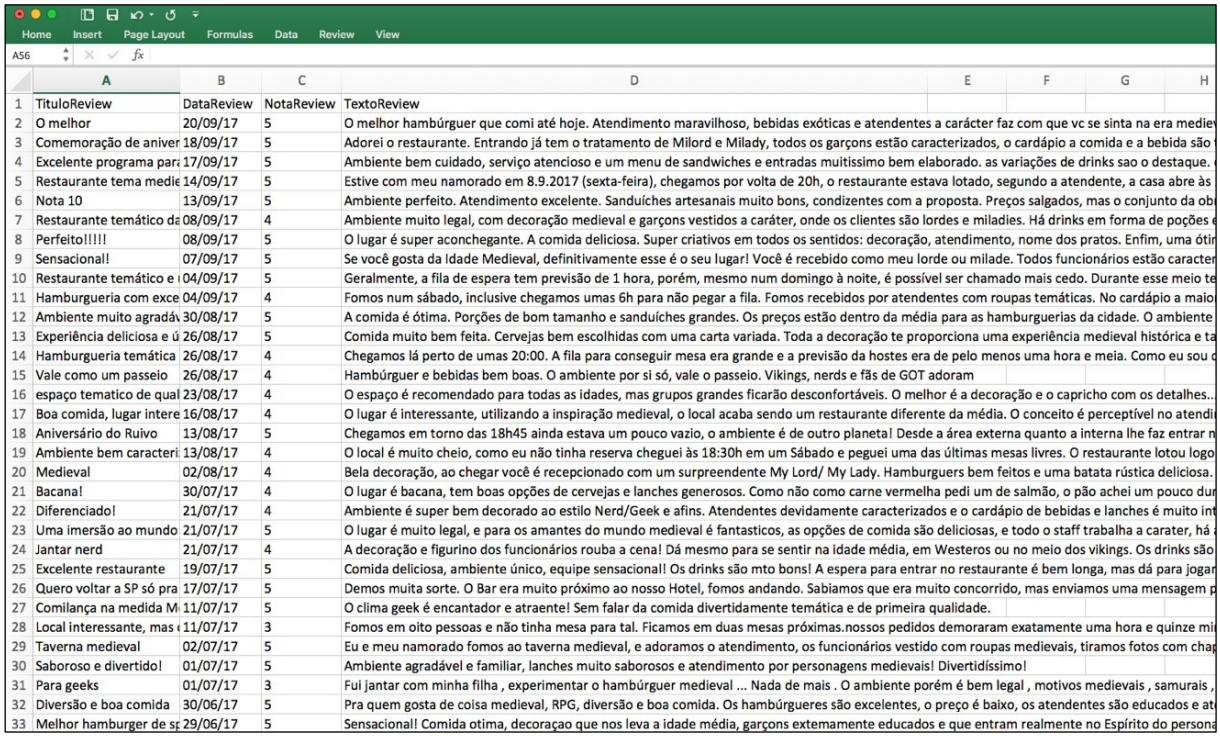
Figura 15 – Escala de notas de uma opinião de usuário do TripAdvisor

Fonte: Elaborado pelo autor.

A conversão da informação da nota foi realizada usando-se substituição simples, fazendo com que o valor original *“ui_bubble_rating bubble_40”* fosse transformado em *‘4’*, o que possibilitará realizar análises envolvendo dados temporais e notas de usuários. Ao final da

primeira fase do pré-processamento, obteve-se uma tabela de dados devidamente organizada e pronta para ser utilizada na segunda fase de pré-processamento, como mostra a Figura 16.

Figura 16 – Dados após a fase de limpeza e estruturação



	A	B	C	D
1	TituloReview	DataReview	NotaReview	TextoReview
2	O melhor	20/09/17	5	O melhor hambúrguer que comi até hoje. Atendimento maravilhoso, bebidas exóticas e atendentes a carácter faz com que vc se sinta na era mediev
3	Comemoração de aniver	18/09/17	5	Adorei o restaurante. Entrando já tem o tratamento de Milord e Milady, todos os garçons estão caracterizados, o cardápio a comida e a bebida são
4	Excelente programa para	17/09/17	5	Ambiente bem cuidado, serviço atencioso e um menu de sandwiches e entradas muitissimo bem elaborado. as variações de drinks sao o destaque.
5	Restaurante tema medie	14/09/17	5	Estive com meu namorado em 8.9.2017 (sexta-feira), chegamos por volta de 20h, o restaurante estava lotado, segundo a atendente, a casa abre às
6	Nota 10	13/09/17	5	Ambiente perfeito. Atendimento excelente. Sanduíches artesanais muito bons, condizentes com a proposta. Preços salgados, mas o conjunto da ob
7	Restaurante temático da	08/09/17	4	Ambiente muito legal, com decoração medieval e garçons vestidos a caráter, onde os clientes são lordes e miladies. Há drinks em forma de poções e
8	Perfeito!!!!!	08/09/17	5	O lugar é super aconchegante. A comida deliciosa. Super criativos em todos os sentidos: decoração, atendimento, nome dos pratos. Enfim, uma ótir
9	Sensacional!	07/09/17	5	Se você gosta da Idade Medieval, definitivamente esse é o seu lugar! Você é recebido como meu lorde ou milade. Todos funcionários estão caracter
10	Restaurante temático e	04/09/17	5	Geralmente, a fila de espera tem previsão de 1 hora, porém, mesmo num domingo à noite, é possível ser chamado mais cedo. Durante esse meio te
11	Hamburgueria com exce	04/09/17	4	Fomos num sábado, inclusive chegamos umas 6h para não pegar a fila. Fomos recebidos por atendentes com roupas temáticas. No cardápio a maio
12	Ambiente muito agradáv	30/08/17	5	A comida é ótima. Porções de bom tamanho e sanduíches grandes. Os preços estão dentro da média para as hamburguerias da cidade. O ambiente
13	Experiência deliciosa e ú	26/08/17	5	Comida muito bem feita. Cervejas bem escolhidas com uma carta variada. Toda a decoração te proporciona uma experiência medieval histórica e ta
14	Hamburgueria temática	26/08/17	4	Chegamos lá perto de umas 20:00. A fila para conseguir mesa era grande e a previsão da hostes era de pelo menos uma hora e meia. Como eu sou
15	Vale como um passeio	26/08/17	4	Hambúrguer e bebidas bem boas. O ambiente por si só, vale o passeio. Vikings, nerds e fãs de GOT adoram
16	espaço tematico de qual	23/08/17	4	O espaço é recomendado para todas as idades, mas grupos grandes ficarão desconfortáveis. O melhor é a decoração e o capricho com os detalhes..
17	Boa comida, lugar intere	16/08/17	4	O lugar é interessante, utilizando a inspiração medieval, o local acaba sendo um restaurante diferente da média. O conceito é perceptível no atendi
18	Aniversário do Ruivo	13/08/17	5	Chegamos em torno das 18h45 ainda estava um pouco vazio, o ambiente é de outro planeta! Desde a área externa quanto a interna lhe faz entrar n
19	Ambiente bem caracteri	13/08/17	4	O local é muito cheio, como eu não tinha reserva cheguei às 18:30h em um Sábado e peguei uma das últimas mesas livres. O restaurante lotou logo
20	Medieval	02/08/17	4	Bela decoração, ao chegar você é recepcionado com um surpreendente My Lord/ My Lady. Hamburguers bem feitos e uma batata rústica deliciosa.
21	Bacana!	30/07/17	4	O lugar é bacana, tem boas opções de cervejas e lanches generosos. Como não como carne vermelha pedi um de salmão, o pão achei um pouco dur
22	Diferenciado!	21/07/17	4	Ambiente é super bem decorado ao estilo Nerd/Geek e afins. Atendentes devidamente caracterizados e o cardápio de bebidas e lanches é muito int
23	Uma imersão ao mundo	21/07/17	5	O lugar é muito legal, e para os amantes do mundo medieval é fantasticos, as opções de comida são deliciosas, e todo o staff trabalha a carater, há
24	Jantar nerd	21/07/17	4	A decoração e figurino dos funcionários rouba a cena! Dá mesmo para se sentir na idade média, em Westeros ou no meio dos vikings. Os drinks são
25	Excelente restaurante	19/07/17	5	Comida deliciosa, ambiente único, equipe sensacional! Os drinks são mto bons! A espera para entrar no restaurante é bem longa, mas dá para jogar
26	Quero voltar a SP só pra	17/07/17	5	Demos muita sorte. O Bar era muito próximo ao nosso Hotel, fomos andando. Sabiamos que era muito concorrido, mas enviamos uma mensagem p
27	Comilança na medida M	11/07/17	5	O clima geek é encantador e atraente! Sem falar da comida divertidamente temática e de primeira qualidade.
28	Local interessante, mas	11/07/17	3	Fomos em oito pessoas e não tinha mesa para tal. Ficamos em duas mesas próximas.nossos pedidos demoraram exatamente uma hora e quinze mi
29	Taverna medieval	02/07/17	5	Eu e meu namorado fomos ao taverna medieval, e adoramos o atendimento, os funcionários vestido com roupas medievais, tiramos fotos com cha
30	Saboroso e divertido!	01/07/17	5	Ambiente agradável e familiar, lanches muito saborosos e atendimento por personagens medievais! Divertidíssimo!
31	Para geeks	01/07/17	3	Fui jantar com minha filha , experimentar o hambúrguer medieval ... Nada de mais . O ambiente porém é bem legal , motivos medievais , samurais ,
32	Diversão e boa comida	30/06/17	5	Pra quem gosta de coisa medieval, RPG, diversão e boa comida. Os hambúrgueres são excelentes, o preço é baixo, os atendentes são educados e at
33	Melhor hamburger de sp	29/06/17	5	Sensacional! Comida otima, decoraçao que nos leva a idade média, garçons extemamente educados e que entram realmente no Espírito do persona

Fonte: Elaborado pelo autor.

A segunda etapa de pré-processamento, aqui denominada de 'Normalização dos dados', deu-se já no ambiente da linguagem R, assim como as demais etapas, incluindo-se a análise de sentimentos, a modelagem de tópicos e a geração de visualizações.

Todos esses processos citados têm como ponto de partida a coluna *TextoReview*, que contém as opiniões criadas pelos clientes dos restaurantes, configurando-se no principal objeto desta pesquisa. É importante ressaltar que esta coluna não recebeu nenhum tratamento prévio na etapa anterior realizada no Microsoft Excel, o que garante a originalidade destes documentos tal qual foram extraídos do site TripAdvisor Brasil.

O primeiro passo para utilizar a linguagem R para as análises foi instalar os pacotes (*packages*) necessários no ambiente local no qual as análises ocorreram. Isso é realizado usando o comando *'install.packages'* (*r-core*) ou a própria interface do RStudio. Após a instalação dos pacotes, é necessário carregá-los no início de cada seção, o que é feito por meio do comando *'library'* (*r-core*). O Quadro 10 mostra o código para carregamento de todos os pacotes necessários ao desenvolvimento das etapas desta pesquisa.

Quadro 10 – Código para carregamento dos pacotes necessários

```
# Carregando pacotes necessários.
library(tidytext)
library(tidyverse)
library(reshape2)
library(readxl)
library(stringr)
library(dplyr)
library(widyr)
library(ptstem)
library(ggplot2)
library(wordcloud)
library(igraph)
library(ggraph)
library(topicmodels)
library(scales)
```

Fonte: Elaborado pelo autor

Uma vez carregados os pacotes, é possível iniciar as análises. Assim, o segundo passo é carregar os dados que serão utilizados na análise, bem como a lista de *stop words* que será usada e ainda os léxicos de polaridade, que serão aplicados no processo de análise de sentimentos das opiniões coletadas. O Quadro 11 mostra o código para carregamento dos dados, léxico e lista de *stop words* desta pesquisa.

Quadro 11 – Código para carregamento dos dados, léxico e lista de *stop words*

```
#################################################################
# Carregando dados, léxicos e stop words
#################################################################

# Carregando léxico 'oplexicon'
lexico <-
  read_excel("~/Dados/oplexicon_v3_0.xlsx") %>%

  # Removendo colunas desnecessárias para as análises
  select(-type,-polarity_revision) %>%

  # Renomeando coluna de polaridade
  rename(polaridade = polarity)

# Carregando lista customizada de stop words
stopwords_pt <-
  read_excel("~/Dados/stopwords_pt.xlsx")

# Carregando dados
reviews <-
  read_excel(
    "~/Dados/TripAdvisor_EXPERIMENTO_PRELIMINAR.xlsx",
    sheet = "DadosTratados",
    col_types = c("text",
                  "date", "numeric", "text")
  )
```

Fonte: Elaborado pelo autor

Após a carga dos dados, o processo ‘Normalização dos dados’ (segunda parte da fase de pré-processamento), compreende a exclusão da coluna *‘TituloReview’*, que não será usada nas análises, a criação de um índice para cada documento, a geração de *tokens*, a conversão de maiúsculas para minúsculas e a remoção de *stop words*, números e caracteres especiais, como *emojis*., conforme código apresentado no Quadro 12.

Quadro 12 – Código correspondente ao processo 'Normalização dos dados'

```
# Carregando dados
reviews_tokens <- reviews %>%

  # Excluindo a coluna 'TituloReview'
  select(-TituloReview) %>%

  # Criando um ID para cada documento
  mutate (document_id = row_number()) %>%

  # Criando tokens por documento
  unnest_tokens(word, TextoReview, to_lower = FALSE, drop = FALSE) %>%

  # Posicionando a coluna 'document_id' no começo
  select(document_id, everything()) %>%

  # Convertendo caracteres em minúsculas
  mutate(word = str_to_lower(word)) %>%

  # Removendo stop words com 'anti_join'
  anti_join(stopwords_pt, by = c("word" = "word")) %>%

  # Filtrando caracteres especiais e números
  filter(str_detect(word, "[a-z]")) %>%
```

Fonte: Elaborado pelo autor

É importante ressaltar que, para efeito didático, o passo no qual ocorre a transformação de maiúsculas para minúsculas poderia ter sido realizada automaticamente durante a criação dos *tokens*, alterando-se a propriedade 'to_lower = FALSE', o que daqui em diante será feito como tal. Importante ressaltar também que o processo de criação de *tokens* elimina automaticamente pontuações.

A remoção de *stop words* ocorreu, neste caso, tal qual explicado anteriormente: por meio de uma operação que compara o conteúdo de duas colunas e, como resultado, mantém apenas na coluna 'B' aquilo que não encontrou na coluna 'A' (comando 'anti_join'). Logo em seguida, os resultados são filtrados por meio de um comando 'filter', eliminando-se números e caracteres especiais.

Entretanto, a exemplo da operação de conversão de maiúsculas para minúsculas, ambos os comandos poderiam ser simplificados e unidos em uma única operação, como exibido no Quadro 13.

Quadro 13 – Código que remove *stop words*, números e caracteres especiais

```
# Removendo stop words e filtrando caracteres especiais e números
  filter(!word %in% stopwords_pt$word, str_detect(word, "[a-z]")) %>%
```

Fonte: Elaborado pelo autor

As *stop words* usadas neste experimento foram retiradas do pacote *tm*, criado por Ingo Feinerer e Kurt Hornik (FEINERER; HORNIK, 2017) e muito popular em trabalhos de mineração de textos que usam a linguagem R como base. Nesta pesquisa, sua única aplicação foi a geração da lista inicial de *stop words*, que posteriormente foi acrescida de dois termos (‘q’ e ‘vc’), dado que ambos apareciam com elevada frequência nas análises iniciais, o que, por sua vez, interferiu nos resultados preliminares. A lista com as *stop words* finais consideradas para este experimento é disponibilizada no Apêndice B.

Há de se destacar um detalhe em relação ao pré-processamento que tem a ver com a fase de redução dos termos ao seu radical, fase comumente aplicada em processos de mineração de texto. Esta etapa, conhecida na literatura como *stemming*, funciona como uma espécie de redução de dimensionalidade (FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2017). Nesta pesquisa, o processo de *stemming* foi realizado, mas os processos de mineração foram realizados sobre os termos completos por, basicamente dois motivos. O primeiro é que não se pretende diminuir a dimensionalidade dos textos utilizados nesta pesquisa, dado que já se trabalha com a limitação destes serem relativamente pequenos, comparando-se aos textos empregados em outras pesquisas, que costumam usar uma quantidade muito maior de documentos. O segundo motivo tem a ver com o problema da eficiência dos *stemmers* para a língua portuguesa, considerando-se sua complexidade.

Durante a realização desta pesquisa, foram testados três algoritmos disponíveis no pacote *ptstem* (FALBEL, 2017): o algoritmo de Porter, o Hunspell e o RSLP. Mesmo considerando sua não utilização nesta pesquisa, apresenta-se no Quadro 14 o código que gerou a variável com os termos reduzidos ao radical, bem como os parâmetros necessários para a execução do processo de *stemming*.

Quadro 14 – Código responsável pela geração do *stemming* aplicado à coluna ‘word’

```
# Stemming
  mutate(stem = ptstem(word, algorithm = "rslp", complete = FALSE)) %>%
```

Fonte: Elaborado pelo autor

Para fins de sumarização dos processos de pré-processamento utilizados nesta pesquisa, o Quadro 15 os apresenta resumidamente.

Quadro 15 – Sumarização dos processos de pré-processamento

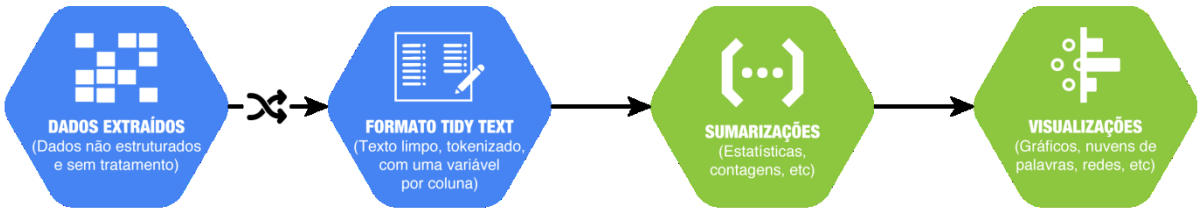
Código	Aplicação
`reviews_tokens <- reviews %>%`	Carrega os dados
`select(-TituloReview) %>%`	Exclui a coluna 'TituloReview' do *data.set*
`mutate (document_id = row_number()) %>%`	Cria um ID para cada documento
`unnest_tokens(word, TextoReview, to_lower = FALSE, drop = FALSE) %>%`	Cria *tokens* por documento
`select(document_id, everything()) %>%`	Posiciona a coluna 'document_id' no começo
`mutate(word = str_to_lower(word)) %>%`	Converte caracteres em minúsculas
`anti_join(stopwords_pt, by = c("word" = "word")) %>%`	Remove *stop words* com 'anti_join'
`filter(str_detect(word, "[a-z]")) %>%`	Remove caracteres especiais e números
`mutate(stem = ptstem(word, algorithm = "rslp", complete = FALSE)) %>%`	Executa *stemming*

Fonte: Elaborado pelo autor.

Esta etapa encerra a fase de pré-processamento dos documentos coletados. Daqui em diante as operações são relacionadas às próximas etapas presentes na *Framework* para Mineração de Opiniões apresentado no início deste capítulo, incluindo a Análise de Sentimentos, Modelagem de Tópicos, Sumarizações e Visualizações.

Dado que o *framework* apresentado no início deste capítulo apoia-se em tarefas clássicas de mineração de textos (FELDMAN; SANGER, 2007), apresenta-se na Figura 17 o esquema de mineração de textos tal qual aplicado nesta pesquisa.

Figura 17 – Esquema da Mineração de Textos aplicado nesta pesquisa



Fonte: Elaborado pelo autor.

Antes de proceder às demais etapas propostas no *Framework* para Mineração de Opiniões aplicado nesta pesquisa, convém descrever melhor os dados da empresa usada como EXPERIMENTO PRELIMINAR. Assim, considerando-se a entrada de dados e o resultado da massa de textos após cada etapa do pré-processamento, pretende-se obter-se uma melhor visão dos dados sobre os quais foram realizadas as etapas de Análise de Sentimentos e Modelagem de Tópicos, conforme indicado no Quadro 16.

Quadro 16 – Resultado da massa de dados após execução dos procedimentos de pré-processamento do EXPERIMENTO PRELIMINAR

Procedimento	Quantidade de itens após procedimento
Entrada	553 documentos
Tokenização	36.269 palavras
Remoção de *stop words*	20.305 palavras
Remoção de números e caracteres especiais	19.800 palavras

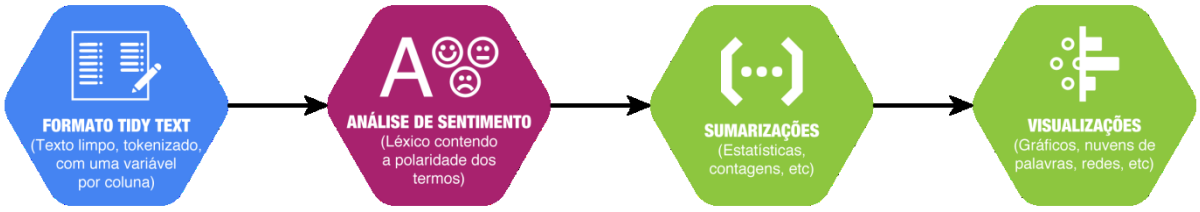
Fonte: Elaborado pelo autor

Como resultado do pré-processamento obteve-se uma população de 19.800 palavras sem *stop words*, distribuídas em 553 documentos estruturados. Estes termos não são únicos, ou seja, repetem-se por todo o *corpus* considerado no EXPERIMENTO PRELIMINAR, o que permite verificar a frequência de sua repetição e o uso deste cálculo para diferentes operações a serem realizadas nas próximas etapas, como a Análise de Sentimentos, Modelagem de Tópicos e visualizações de termos relevantes.

Assim, o próximo passo foi a Análise de Sentimentos, que nesta pesquisa foi realizada usando-se léxicos. A abordagem baseada em léxicos é melhor explicada no tópico ‘Análise de Sentimento’, bem como o motivo de sua escolha para a realização desta pesquisa.

A Análise de Sentimentos tal qual aplicada neste trabalho é apresentada na Figura 18. O processo de Análise de Sentimentos com a abordagem de dados arrumados é detalhado mais adiante.

Figura 18 – Esquema da Análise de Sentimentos aplicada nesta pesquisa



Fonte: Elaborado pelo autor.

Conforme explicado anteriormente, a análise de sentimentos (usando a abordagem de *tidy data*) pode ser realizada por meio de uma operação tipo '*inner_join'* entre duas colunas de dados, uma contendo os *tokens* gerados na fase de pré-processamento e outra contendo os termos do léxico.

O cálculo de quão positiva ou negativa uma opinião é classificada, baseia-se na soma das polaridades dos termos contidos na opinião e constantes do léxico (PANG; LEE, 2008; LIU, 2012; SILGE; ROBINSON, 2017). Esta não é a única maneira de abordar a análise do sentimento de conteúdos textuais, mas é frequentemente usada. Esta abordagem tem suas limitações, sobretudo por conta de a análise ser feita especificamente em relação a uma unidade de observação (um *token*).

Como explicado no tópico 'Análise de Sentimento', a escolha pelo método léxico deve-se sobretudo às limitações em relação ao universo de dados abordados. Outro ponto que deve ser observado é a escassez de recursos léxicos validados em português do Brasil. Para esta pesquisa e foram experimentados os dois léxicos mais comumente utilizados disponíveis na literatura em português do Brasil: opLexicon versão 3.0 (SOUZA; VIEIRA, 2012) e o sentiLex (SILVA *et al.*, 2010).

O léxico opLexicon possui 32.191 termos classificados em três categorias: termos positivos (polaridade 1), termos negativos (polaridade -1) e termos neutros (polaridade 0). O sentiLex possui 7.014 termos classificados nas mesmas categorias que o opLexicon. Além das polaridades, estes léxicos possuem outras variáveis, como o tipo gramatical do termo e outras que não foram usadas nesta pesquisa. O quadro 4 mostra a quantidade de itens positivos, negativos e neutros em cada léxico citado.

Quadro 17 – Quantidade de termos por polaridade no opLexicon e no sentiLex

Polaridade	opLexicon	sentiLex
Termos positivos	8.620	1.548
Termos negativos	14.569	4.602
Termos neutros	9.002	860

Fonte: Elaborado pelo autor

Nota-se que ambos os léxicos possuem mais termos negativos que positivos. Devido à abordagem de *tidy data* empregada nesta pesquisa, a operação de análise de sentimentos pode ser realizada, como explicado anteriormente, por meio de um comando '*inner_join'*, conforme exposto no código apresentado no Quadro 18.

Quadro 18 – Código que realiza a Análise de Sentimentos com um comando *'inner_join'*

```
# Cruzando léxico de polaridade com um 'inner_join'
inner_join(oplexicon_v3_0, by = c("word" = "term")) %>%
```

Fonte: Elaborado pelo autor

O cálculo do quão positiva ou negativa é uma opinião é expresso por meio de uma soma dos termos constantes em cada documento, considerando-se o índice *'document_id'* criado no momento da tokenização, conforme o seguinte código:

Quadro 19 – Código que realiza a soma das polaridades dos termos de um documento

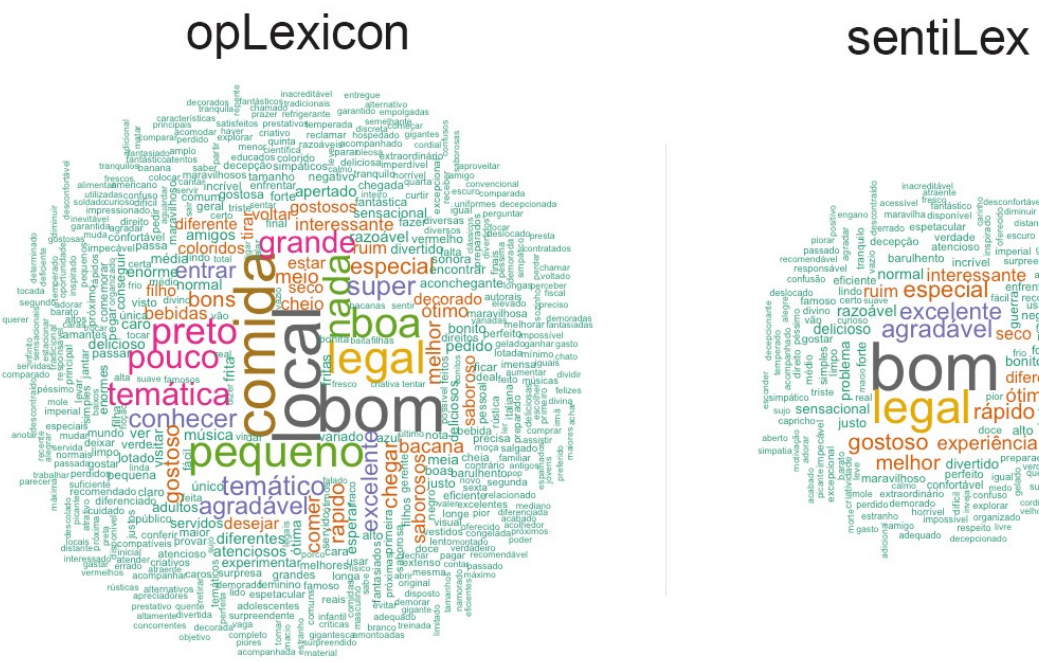
```
# Soma das polaridades
group_by(document_id) %>% mutate(soma_polaridade = sum(polaridade)) %>%
ungroup() %>%
```

Fonte: Elaborado pelo autor

A partir deste ponto, diversas visualizações foram geradas buscando compreender melhor a massa de opiniões geradas pelos clientes dos restaurantes analisados na pesquisa. Em se tratando da análise de sentimentos, o impacto da escolha do léxico evidenciou-se por meio de visualizações consideradas. Elas são apresentadas a seguir, com o intuito de ilustrar e embasar algumas decisões tomadas pelo pesquisador.

Para efeito de rápida comparação entre os resultados obtidos ao se usar ambos os léxicos, apresenta-se a seguir algumas visualizações. A primeira visualização é a nuvem de termos mais comuns, gerada a partir da frequência dos termos classificados pelos léxicos considerados, conforme exposto na Figura 19.

Figura 19 – Nuvem de termos mais frequentes após cruzamento com os léxicos[3]



Fonte: Elaborado pelo autor.

Observando-se ambas as nuvens geradas após o cruzamento da massa de dados com os dois léxicos aplicados nesta pesquisa, algumas características chamam a atenção, a começar pela relação entre o tamanho de ambas, que se explica pelo respectivo tamanho dos léxicos empregados no experimento, conforme visto no Quadro 17.

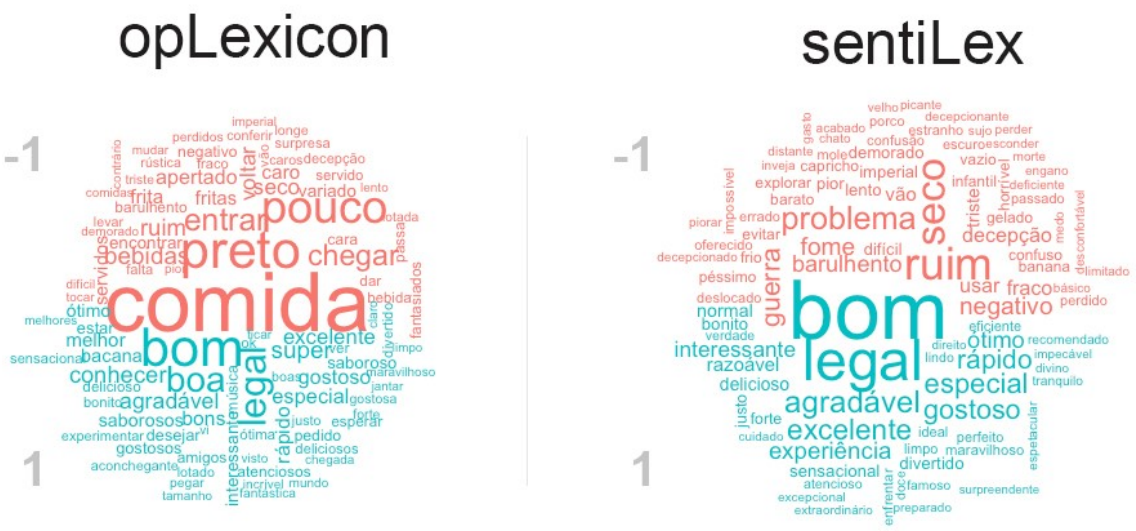
Isto deve-se ao fato do tamanho do léxico opLexicon ser bem maior que o léxico sentiLex. Outra curiosidade volta-se aos termos que aparecem nas análises que empregam o opLexicon, mas não das análises com o léxico sentiLex. Os termos ‘local’ e ‘comida’ não constam em ambos os léxicos, o que leva a crer que, a depender do aspecto analisado, o opLexicon será capaz de proporcionar um resultado melhor, ao menos para esta pesquisa.

Por outro lado, o opLexicon avalia certos termos como negativos, que no contexto do *corpus* analisado, deveriam ser, no mínimo, neutros. Um exemplo disso é o termo ‘comida’, que no léxico opLexicon é negativo, e no léxico sentiLex sequer consta como um termo, como mostra a Figura 20.

[3] Os tamanhos das nuvens foram mantidos tal qual foram gerados pelo R a fim de evidenciar a frequência de termos após o cruzamento com os léxicos.

Outro indício sobre o impacto da escolha do léxico nos resultados das análises pode ser observado na nuvem de termos positivos e negativos de ambos os léxicos em relação ao *corpus*, conforme Figura 20.

Figura 20 – Nuvem de termos positivos e negativos, segundo os léxicos empregados



Fonte: Elaborado pelo autor.

Outra ação tomada após o cruzamento dos dados dos dois léxicos com a massa de dados obtida depois do pré-processamento, foi a criação de uma representação vetorial dos termos restantes no *corpus* considerado, o que se deu por meio do cálculo de pesos dos termos em relação à sua repetição no *corpus*. Esta representação, conhecida como TF-IDF (da abreviação do inglês *Term Frequency–Inverse Document Frequency*) significa a aferição da frequência do termo–inverso em relação à sua frequência nos documentos. Trata-se de uma medida estatística que tem o intuito de indicar a importância da palavra de um documento em relação a uma coleção de documentos ou em um *corpus* linguístico.

A ideia do uso da representação TF-IDF é encontrar as palavras importantes para o conteúdo de cada documento, diminuindo o peso das palavras comumente usadas e aumentando o peso das palavras que não são usadas com frequência ampliada numa coleção ou *corpus* de documentos, neste caso, as opiniões dos usuários do restaurante como um todo. O cálculo do TF-IDF busca encontrar as palavras importantes (ou seja, comuns) em um texto, mas não muito comuns. Neste experimento a representação TF-IDF foi implementada com o código apresentado no Quadro 20.

Quadro 20 – Código que conta a quantidade de termos por documento e calcula o TF-IDF

```
# Contando a quantidade de repetições de termos por documento
group_by(document_id, word) %>% mutate(n = n()) %>% ungroup() %>%

# Calculando TF-IDF
bind_tf_idf(word, document_id, n) %>%
```

Fonte: Elaborado pelo autor

A função 'bind_tf_idf' considera três parâmetros: a coluna contendo os *tokens*, a coluna contendo o índice de cada documento e o número de vezes que um termo aparece num mesmo documento. O resultado é a criação de três colunas contendo o valor das frequências direta e inversa de cada termo no *corpus* analisado, e ainda o valor TF-IDF usado para representar a importância de um determinado termo no *corpus* em análise.

O resultado das análises de sentimento envolvendo os dois léxicos considerados nesta pesquisa apresentaram comportamentos diferentes, dado que eles variam tanto em tamanho quanto na concentração de termos com polaridade positiva, negativa e neutra.

Tendo em mente as comparações apresentadas, optou-se pela aplicação de ambos os léxicos no restante da pesquisa visando compara-los, mesmo considerando-se que nenhum dos dois atende totalmente ao domínio abordado neste trabalho.

Considerando-se a limitação de ambos os léxicos, outra análise que poderia contribuir na descoberta de conhecimento tem a ver com a combinação de dois ou mais termos, o que é conhecido na literatura como *n-grams* (ou n-gramas), que podem ser formados pela concatenação de dois ou mais *tokens*.

Um n-grama é uma sequência de n itens dentro de uma frase, e podem ser constituídos de palavras, letras, sílabas, classificação gramatical das palavras, ou qualquer outra base. Um n-grama de tamanho 1 é chamado de uni-grama, de tamanho 2, bi-grama, de tamanho 3, tri-grama, de 4 ou mais, n-grama.

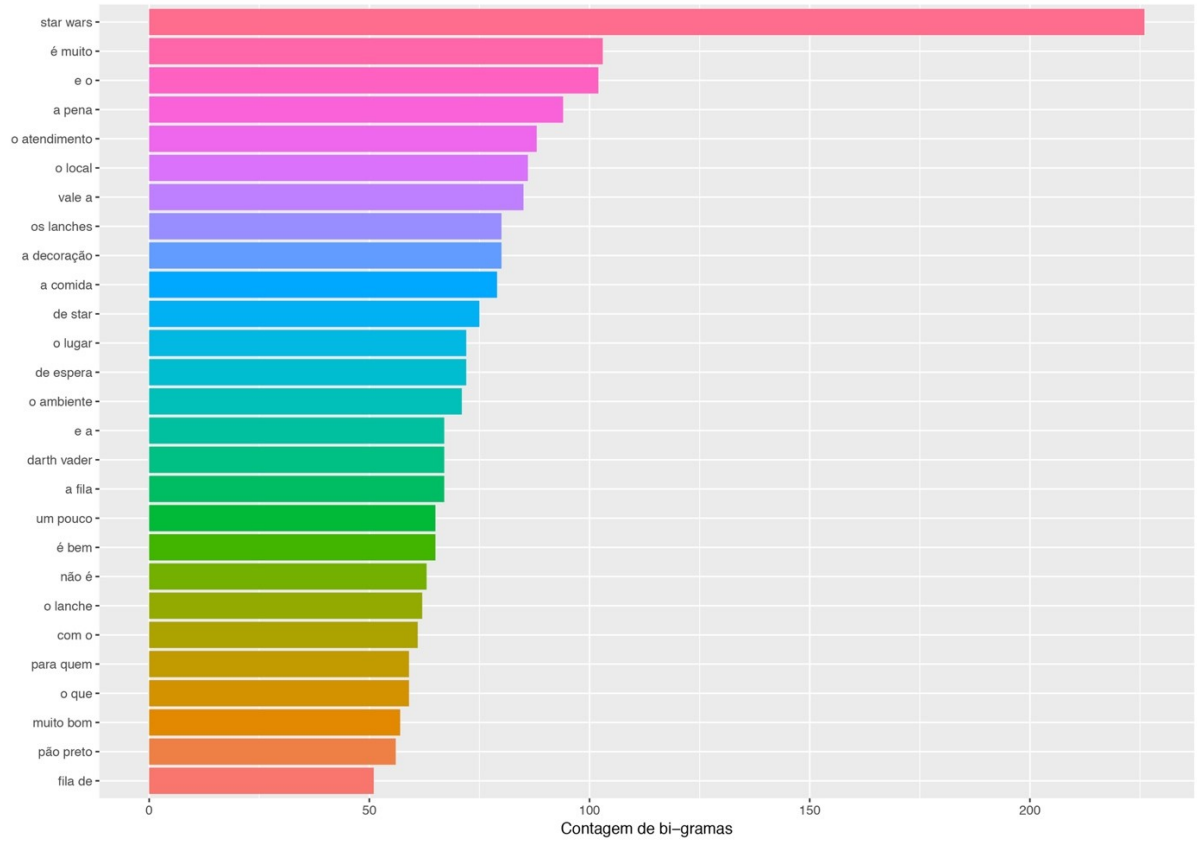
A literatura mostra que a utilização de n-gramas proporciona bons resultados, dado que este é um atributo bastante informativo e cumpre um importante papel para capturar estilos linguísticos de textos com polaridade, inclusive sendo aplicado em trabalhos que consideram múltiplos idiomas (ABBASI; CHEN; SALEM, 2008; ABBASI *et al.*, 2011).

Neste experimento adotou-se a análise de bi-gramas, ou seja, a concatenação de dois *tokens* formados, por sua vez, de palavras que aparecem imediatamente juntas nos documentos (FELDMAN; SANGER, 2007; SILGE; ROBINSON, 2017).

O tratamento dado aos bi-gramas é exatamente o mesmo dispensado aos *tokens* únicos, ou seja, passam pelo mesmo pré-processamento que os *tokens*, com apenas uma exceção particular para este experimento, qual seja: não foram retiradas as *stop words*. A explicação para tal decisão advém do fato de a lista original de *stop words* remover advérbios de negação como ‘não’, ‘nem’, ‘nunca’, além de outros termos como 'muito', que normalmente são usados como intensificadores, a exemplo de ‘muito ruim’ ou ‘muito bom’. Como estes bi-gramas tratam de informação importante para as análises do *corpus* deste experimento, decidiu-se manter tais termos.

Entre várias possibilidades possíveis ao analisar-se bi-gramas, pode-se listar a frequência na qual aparem no *corpus* antes mesmo de proceder à análise de sentimentos, como mostra a Figura 21.

Figura 21 – Bi-gramas mais comuns no *corpus*



Fonte: Elaborado pelo autor.

Analisando-se a Figura 21, percebe-se uma quantidade elevada de termos considerados *stop words*. Entretanto, ao proceder análises específicas, como por exemplo, quais termos são precedidos por outros termos, é possível encontrar resultados interessantes e que

não seriam possíveis caso as *stop words* fossem previamente removidas. No exemplo do Quadro 21, filtram-se os bi-gramas mais frequentes precedidos por termo 'muito', possibilitando assim expor combinações interessantes para a análise das peculiaridades do *corpus* considerado neste trabalho.

Quadro 21 – Código que filtra palavras precedidas pelo termo 'muito'

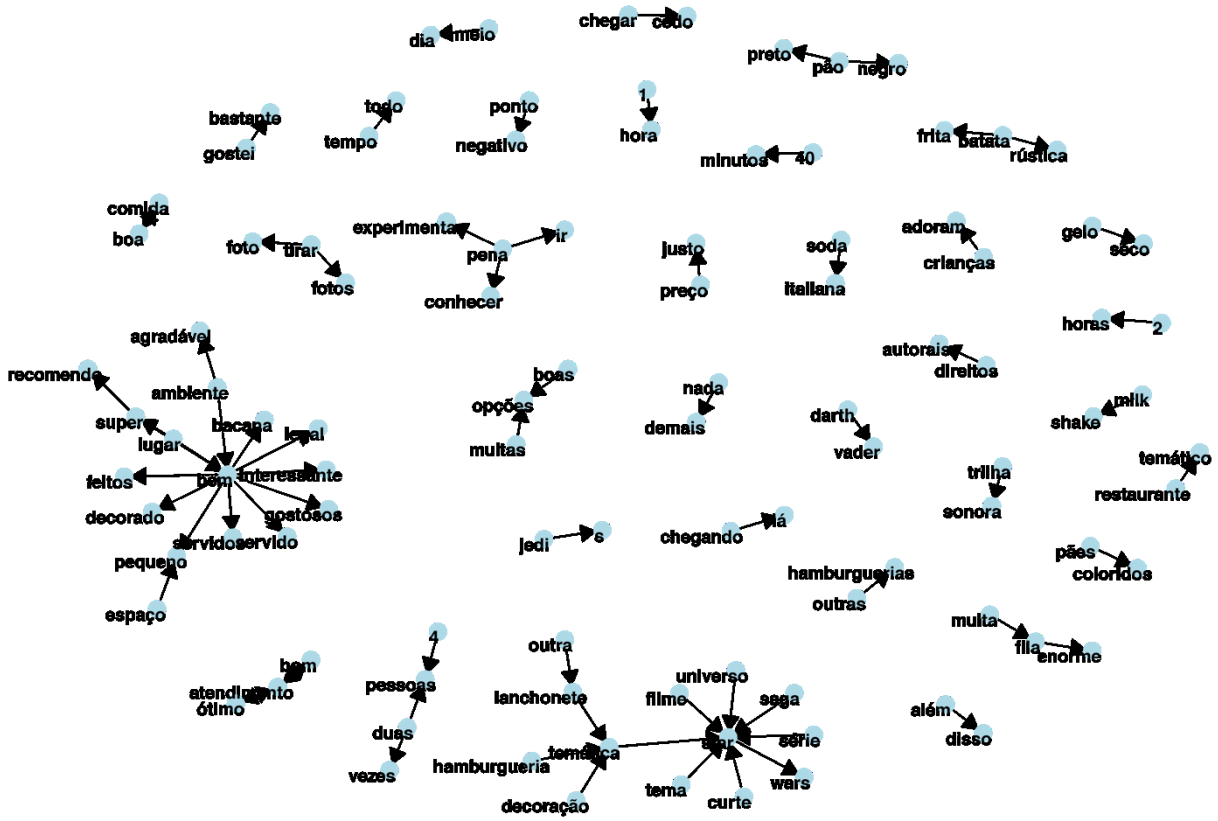
```
# Palavras mais frequentes precedidas por 'muito'
muito_words <- bigramas_separados %>%
  filter(word1 == "muito") %>%
  inner_join(lexico, by = c(word2 = "term")) %>%
  count(word2, polaridade, sort = TRUE) %>%
  ungroup()

# Palavras mais frequentes precedidas por 'muito'
muito_words

## # A tibble: 92 x 3
##        word2 polaridade     n
##        <chr>      <dbl> <int>
##  1       bom          1    57
##  2     legal          1    34
##  3       boa          1    31
##  4   gostoso          1    14
##  5    grande          0    12
##  6   pequeno          0    11
##  7      bons          1    10
##  8 agradável          1     9
##  9 saborosos          1     9
## 10  saboroso          1     8
## # ... with 82 more rows
```

Fonte: Elaborado pelo autor

Caso as *stop words* tivessem sido removidas, seria difícil encontrar resultados como os exibidos no Quadro 21. Ou seja, 'muito pequeno' (que denota conotação negativa) é algo que aparece juntamente com vários termos supostamente positivos, como 'muito saboroso' e 'muito gostoso' (relativos aos pratos servidos no restaurante), o que pode ser um indício de que há algo errado com o tamanho das porções servidas.

Além disso, outra forma de visualizar as relações entre os bi-gramas contidos no *corpus* analisado se dá por meio do emprego de técnicas de visualização de redes de palavras, que nesta pesquisa foram implementadas com apoio dos pacotes *igraph* (CSARDI; NEPUSZ, 2006) e *ggraph* (PEDERSEN, 2017), conforme exposto na Figura 22.

Figura 22 – Visualização da rede de palavras

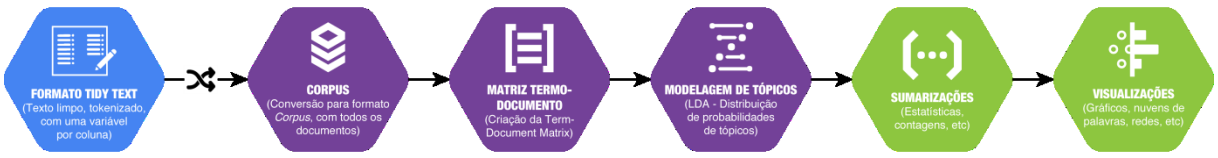


Fonte: Elaborado pelo autor.

Para a geração da rede de palavras foi realizada a exclusão de *stop words*. Entretanto, não houve cruzamento dos termos com nenhum dos léxicos empregados, uma vez que tal prática provocaria a redução drástica da quantidade de bigramas, resultando em uma visualização pobre da rede de palavras.

Para aprofundar as análises (e como última parte do *Framework* para Mineração de Opiniões apresentado no começo deste capítulo), aplicou-se a Modelagem de Tópicos à massa de textos, valendo-se da técnica conhecida como *Latent Dirichlet Allocation* (LDA) (BLEI, 2012). A Modelagem de Tópicos aplicada neste trabalho é apresentada na Figura 23, cuja implementação é detalhada mais adiante.

Figura 23 – Esquema da Modelagem de Tópicos aplicada nesta pesquisa



Fonte: Elaborado pelo autor.

A *Latent Dirichlet Allocation* (LDA) é um dos algoritmos mais comuns para modelagem de tópicos (BLEI; LAFFERTY, 2009; BLEI, 2012). Sem mergulhar na matemática por trás do modelo, sua compreensão é guiada por, basicamente, dois princípios:

1. **Cada documento é uma mistura de tópicos -** cada documento pode conter palavras de vários tópicos em proporções específicas. Por exemplo, em um modelo de dois tópicos, seria possível expressar: "O Documento 1 é 90% de tópico A e 10% de tópico B, enquanto o Documento 2 é 30% de tópico A e 70% de tópico B."
2. **Todo tópico é uma mistura de palavras** – considera-se um modelo de dois tópicos. Por exemplo, no caso de notícias, um tópico para 'política' e outro para 'entretenimento'. As palavras mais comuns no tópico sobre política podem ser 'Presidente', 'Congresso' e 'Governo'; enquanto o tema entretenimento pode ser composto de palavras como 'filmes', 'televisão' e 'ator'. Importante notar que as palavras podem ser compartilhadas entre os tópicos, ou seja, uma palavra como "orçamento" pode aparecer em ambos os tópicos.

A LDA é um método matemático para estimar ambos os princípios expostos ao mesmo tempo, ou seja, buscando assim encontrar a mistura de palavras que está associada a cada tópico, ao mesmo tempo que determina a mistura de tópicos que descreve cada documento. Vale destacar que há uma série de implementações existentes deste algoritmo, porém, neste experimento optou-se pelo uso do pacote *topicmodels* (GRÜN; HORNIK, 2011), que implementa a solução original proposta por David M. Blei (BLEI, 2012).

A implementação do algoritmo LDA exige que os dados estejam em um formato específico: uma Matriz Termo-Documento (matriz DTM, do inglês *Document Term Matrix*) (FELDMAN; SANGER, 2007). Uma Matriz Documento-Termo é uma matriz matemática que descreve a frequência de termos que ocorrem em uma coleção de documentos. Em uma Matriz Documento-Termo, as linhas correspondem aos documentos na coleção e as colunas correspondem aos termos. Nesta pesquisa, a matriz Documento-Termo foi implementada da forma apresentada no Quadro 22.

Quadro 22 – Código que cria uma Matriz Documento-Termo e aplica o LDA à matriz

```
# Convetendo data.frame em formato Matriz Termo-Documento
reviews_dtm <- dataframe_tm %>%
  cast_dtm(document_id, word, n)

# Verificandoo a Matriz Documento-Termo
reviews_dtm

## <<DocumentTermMatrix (documents: 535, terms: 3917)>>
## Non-/sparse entries: 18407/2077188
## Sparsity           : 99%
## Maximal term length: 18
## Weighting          : term frequency (tf)

# Usando a função LDA() para o modelo de tópicos
reviews_lda <- LDA(reviews_dtm, k = 4, control = list(seed = 2017))
```

Fonte: Elaborado pelo autor

Após a geração da Matriz Documento-Termo usou-se a função *'LDA'* para criar um modelo de quatro tópicos (k = 4) que gerou, como resultado, um modelo probabilístico de quatro tópicos, conforme visualizado no Quadro 23.

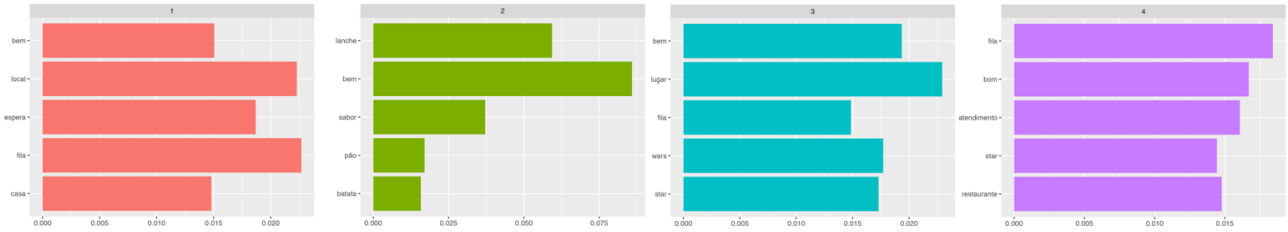
Quadro 23 – Código para visualização dos tópicos gerados pelo modelo LDA

```
# Visualizando os tópicos
top_termos_topicos

## # A tibble: 20 x 3
##    topic        term       beta
##    <int>       <chr>      <dbl>
##  1     1        fila 0.02268653
##  2     1       local 0.02228617
##  3     1      espera 0.01868285
##  4     1         bem 0.01504523
##  5     1        casa 0.01479818
##  6     2         bem 0.08589053
##  7     2      lanche 0.05933485
##  8     2       sabor 0.03714808
##  9     2         pão 0.01702071
## 10     2      batata 0.01577620
## 11     3       lugar 0.02293026
## 12     3         bem 0.01934882
## 13     3        wars 0.01769941
## 14     3        star 0.01729262
## 15     3        fila 0.01484740
## 16     4        fila 0.01841135
## 17     4         bom 0.01669391
## 18     4 atendimento 0.01606219
## 19     4 restaurante 0.01477144
## 20     4        star 0.01442644
```

Fonte: Elaborado pelo autor

O valor ‘beta’ apresentado no Quadro 23 é um valor gerado pelo modelo que corresponde às probabilidades por tópico por palavra, que por sua vez pode ser visualizado graficamente, conforme mostra a Figura 24.

Figura 24 – Distribuição das palavras por tópico



Fonte: Elaborado pelo autor.

Por meio das técnicas apresentadas neste capítulo, demonstrou-se a capacidade do *Framework* para Mineração de Opiniões apresentado no início do capítulo e aplicado nesta

pesquisa de extrair conhecimento das opiniões provenientes dos clientes de restaurantes publicadas na rede social TripAdvisor, escolhida para este trabalho.

Também demonstra sua capacidade de representar o conhecimento de forma a facilitar diferentes tipos de análises das opiniões de cliente dos restaurantes considerados nesta pesquisa.

Cada um dos conjuntos de dados dos três restaurantes restantes (EMPRESA 1, EMPRESA2 e EMPRESA3) será analisado usando o modelo consolidado gerado a partir do EXPERIMENTO PRELIMINAR, e apresentados no capítulo “Apresentação e Análise dos Resultados”, a seguir.

4 APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

Neste capítulo apresenta-se a aplicação do *Framework* para Mineração de Opiniões proposto no capítulo de Procedimentos Metodológicos para os dados das três empresas selecionadas, bem como as análises dos resultados obtidos. Ao final, também apresenta-se a consolidação dos resultados obtidos nas análises efetuadas à luz da literatura da temática abordada.

4.1 Análise dos dados da EMPRESA 1

Assim como as demais empresas selecionadas para a realização desta pesquisa, a EMPRESA 1 é um restaurante especializado em hambúrgueres, do Estado de São Paulo (SP) e está presente no TripAdvisor desde Dezembro de 2009, contando com poucas avaliações à época de sua estreia na rede social. Talvez isto deva-se ao fato de que naquele momento, o TripAdvisor ainda não contasse com uma versão em português do site e não houvesse a mesma divulgação da rede social no Brasil, em comparação com outros países como os Estados Unidos.

O restaurante foi fundado em 1963 e surgiu como uma lanchonete. *Cheese* Salada, *Cheese* Bacon e *Milk Shake* foram alguns dos produtos oferecidos à época de sua inauguração e que estão presentes no cardápio até hoje. O restaurante é bastante popular e ativo em outras redes sociais como o Facebook, com 8.549 seguidores e Instagram, com 1.689 seguidores.

Cabe ressaltar que a EMPRESA 1 possui um Certificado de Excelência concedido pelo TripAdvisor aos estabelecimentos que recebem avaliações excelentes dos usuários com frequência (TRIPADVISOR, 2017).

As análises começam logo após a etapa de pré-processamento de dados, que consiste na criação de *tokens* e remoção de *stop words*, caracteres e números especiais. O Quadro 24 mostra o resultado da massa de dados da EMPRESA 1 após cada etapa de pré-processamento, conforme descrito anteriormente no capítulo que trata dos Procedimentos Metodológicos.

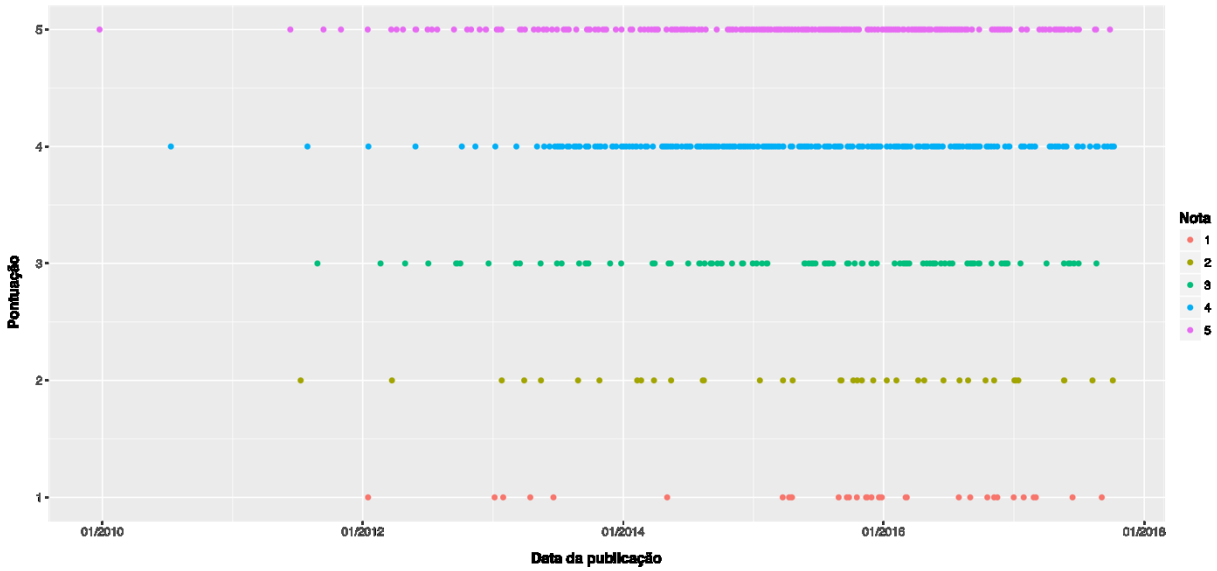
Quadro 24 – Sumarização dos dados da EMPRESA 1 após fase de pré-processamento

Procedimento	Quantidade de itens após procedimento
Entrada	793 documentos
Tokenização	35.243 palavras
Remoção de *stop words*	20.145 palavras
Remoção de números e caracteres especiais	19.808 palavras

Fonte: Elaborado pelo autor

Como resultado das etapas do pré-processamento, obteve-se um conjunto de 19.808 palavras sem *stop words*, distribuídas nos 793 documentos da EMPRESA 1. Interessante notar que este resultado representa apenas oito palavras a mais que o resultado apresentado no Quadro 16, que reflete os dados resultantes do EXPERIMENTO PRELIMINAR. A análise dos dados começa com a visualização da distribuição das notas ao longo do tempo, apresentada na Figura 25.

Figura 25 – Distribuição das notas ao longo do tempo para a EMPRESA 1

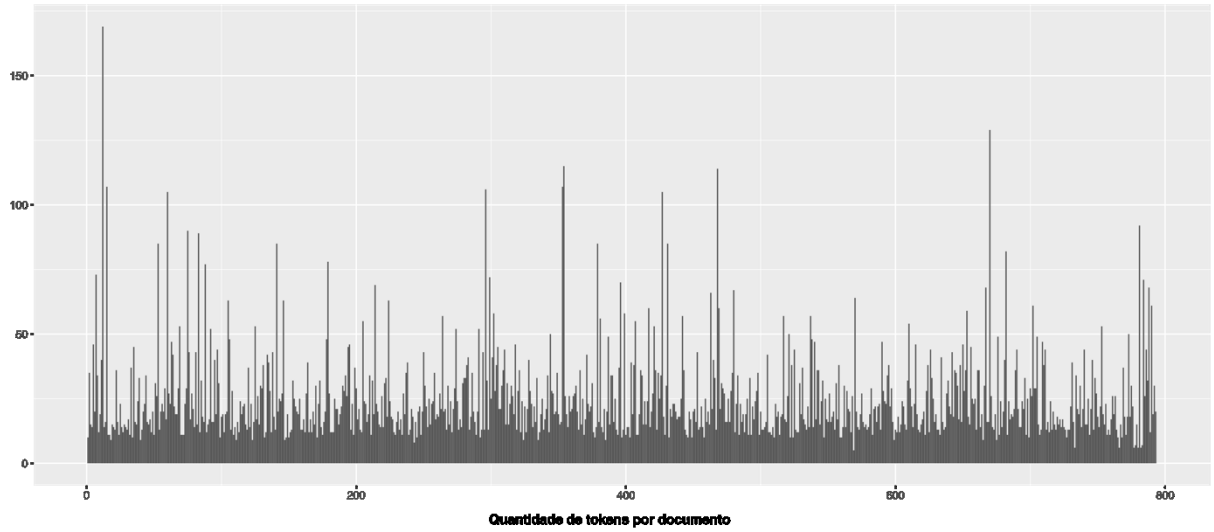


Fonte: Elaborado pelo autor.

Nota-se pela distribuição dos dados no gráfico apresentado na Figura 25 que a empresa possui uma quantidade muito maior de notas positivas do que negativas, e que a última linha (notas 1) possui uma distribuição que permite avaliar que as avaliações negativas não são constantes, ao contrário das avaliações mais positivas (notas 4 e 5).

A próxima análise representa o tamanho médio das avaliações por meio de um histograma que reflete a quantidade de termos por documento, como mostra a Figura 26.

Figura 26 – Quantidade de termos por documento da EMPRESA 1



Fonte: Elaborado pelo autor.

Como pode ser observado no gráfico apresentado na Figura 26, alguns documentos possuem um tamanho longo, enquanto outros são compostos apenas de poucas palavras, o que pode impactar consideravelmente a fase de Análise de Sentimentos, dado que esta é baseada em léxicos e considera os termos positivos e negativos presentes num documento.

Ao analisar a quantidade de repetições de termos no *corpus* em geral, apresentadas no Quadro 25 pelos dez termos mais frequentes, nota-se que alguns destes se destacam e ajudam a corroborar o que foi visto na Figura 25 em relação à distribuição das notas ao longo do tempo, dado que termos mais positivos provavelmente significam uma maior presença de opiniões positivas e, consequentemente, notas mais positivas atribuídas pelos usuários.

Quadro 25 – Dez termos mais presentes em todo o *corpus* da EMPRESA 1

Termo	Quantidade de repetições no *corpus*
atendimento	324
bem	279
bom	266
lanches	243
hambúrguer	186
lanche	186
sempre	170
rápido	166
maionese	165
lugar	152

Fonte: Elaborado pelo autor

Nota-se que, pela repetição absoluta dos termos no *corpus*, os clientes falam muito sobre o atendimento (324 repetições) e sobre a comida (mais de 600 repetições, se considerarmos a soma dos termos lanches, hambúrguer e lanche).

Entretanto, apenas a repetição de termos não representa muito do ponto de vista de informação e conhecimento gerado. Ao comparar-se os dados do Quadro 25 com os dez termos com mais repetições por documento apresentado no Quadro 26, é possível ter uma ideia diferente sobre a frequência de alguns termos.

Quadro 26 – Dez termos com mais repetições por documento da EMPRESA 1

Termo	Quantidade de repetições por documento
chicletes	7
carne	6
hambúrguer	5
hambúrguer	5
parmegiana	5
fritas	5
bacon	5
x	5
mostarda	5
cliente	4

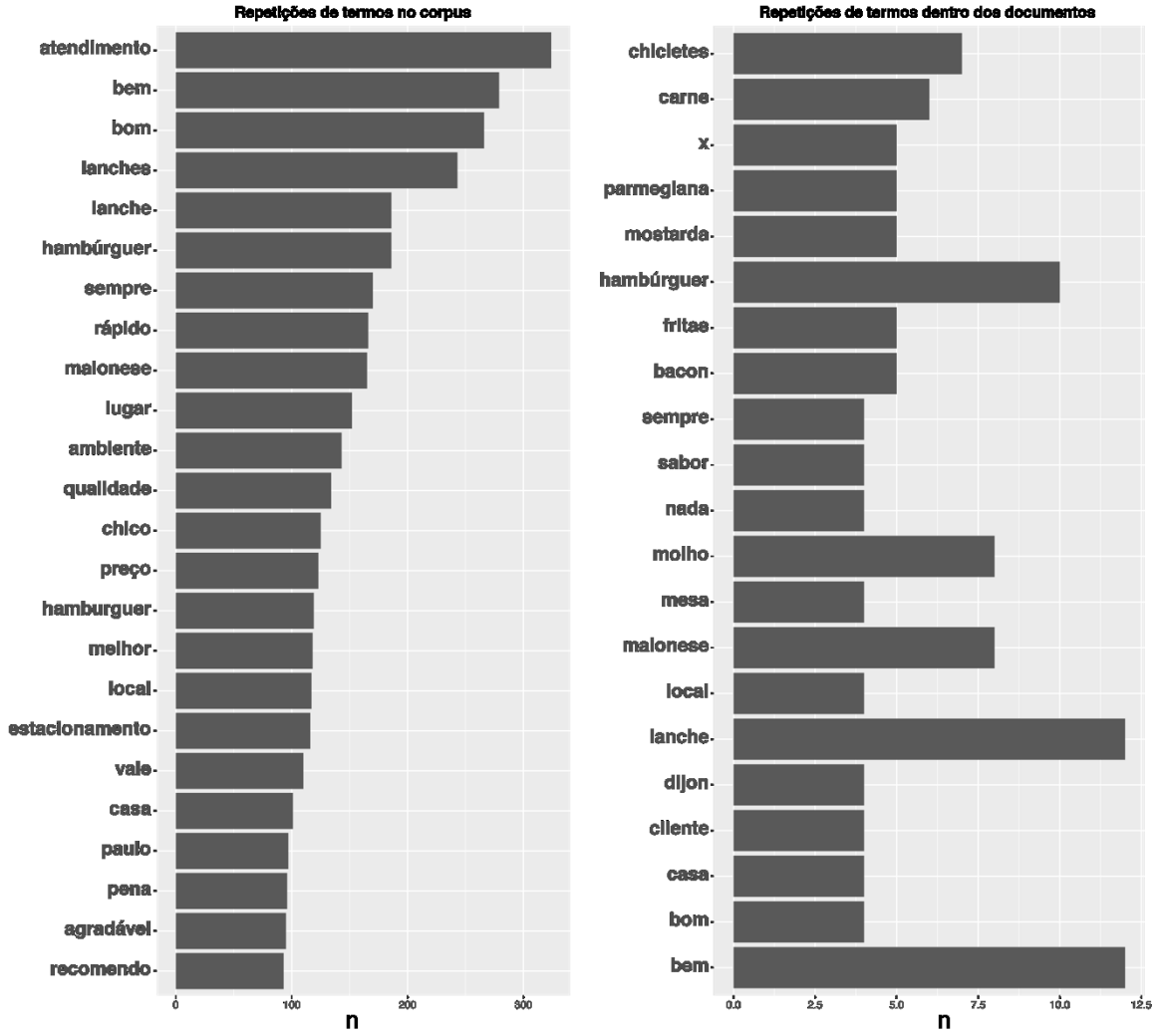
Fonte: Elaborado pelo autor

Surpreendentemente, o termo ‘chicletes’ repete-se sete vezes dentro de algum documento do universo de 793 documentos. Além disso, os dez termos que mais se repetem dentro dos documentos referem-se à comida, especificamente. Nota-se um termo que parece

ser um *outlier*, mas não é: o termo ‘X’, que neste caso refere-se a alguns pratos servidos pelo estabelecimento (x-burguer, x-salada, etc) e passou intacto pela remoção de *stop words*.

No contexto desta pesquisa, um *outlier* é um valor atípico, uma observação que apresenta inconsistência em relação aos demais dados da série. Outra possível análise é o contraste proporcionado pela plotagem dos gráficos de repetição de termos no *corpus* e dentro dos documentos, conforme apresentado na Figura 27.

Figura 27 – Repetição de termos no *corpus* e dentro dos documentos da EMPRESA 1



Fonte: Elaborado pelo autor.

Pela comparação de ambos os gráficos de repetição, verifica-se que alguns termos aparecem em ambos, devido obviamente à sua frequência geral, mas os termos que mais se repetem dentro dos documentos têm a ver com a comida, aspecto importante para esta pesquisa dada a natureza do *corpus* analisado.

Uma outra forma de avaliar a presença de determinados termos em relação ao quanto estes termos se repetem num *corpus* é contrastar termos mais frequentes com termos menos frequentes, o que pode ser feito por meio de uma nuvem de palavras, cujo resultado apresenta-se na Figura 28.

Figura 28 – Nuvem de termos mais frequentes da EMPRESA 1

Fonte: Elaborado pelo autor.

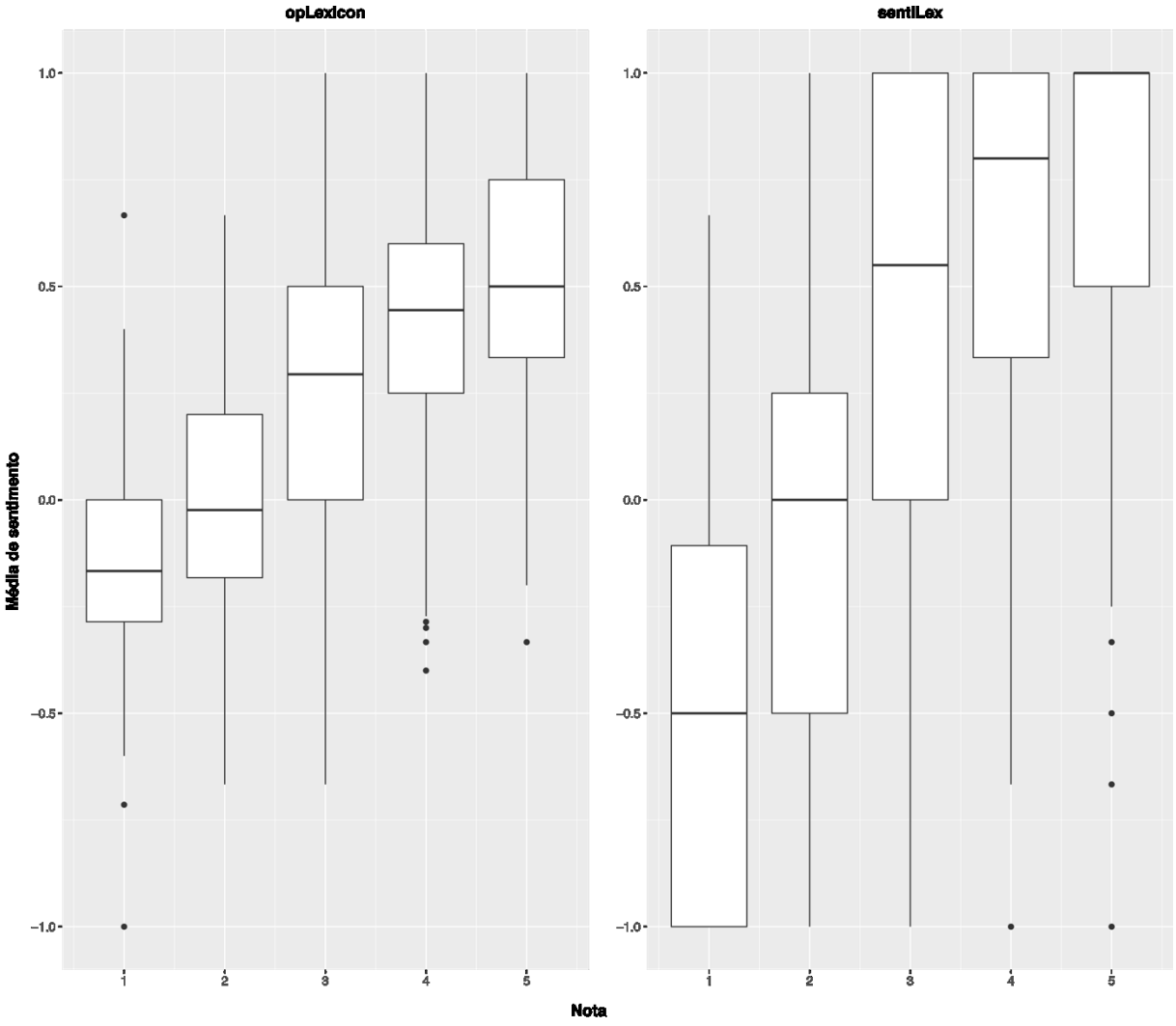
Embora contagens diretas e frequências de termos ajudem a ter uma visão geral da distribuição e regularidade com que as palavras se aparecem, continuam sendo insuficientes para cumprir os objetivos propostos nesta pesquisa. Sendo assim, o próximo passo consiste em usar a Análise de Sentimentos com o objetivo de avaliar as inclinações das postagens dos usuários.

Daqui em diante as análises envolvendo sentimentos serão feitas aplicando-se ambos os léxicos (opLexicon e sentiLex) apresentados no Capítulo 3 - Procedimentos Metodológicos, cujos resultados serão colocados sempre lado a lado, para efeito de comparação. Assim, ao cruzar os léxicos com os dados que passaram pelo pré-processamento,

cada léxico foi capaz de classificar uma quantidade de termos presentes no *corpus*: o opLexicon classificou 5.837 termos e o sentiLex classificou 2.097 termos.

A primeira análise cruza os dados de pontuação atribuída pelo usuário numa escala de 1 a 5 e a classificação de sentimentos atribuída pelo léxico. A Figura 29 mostra um gráfico do tipo *box plot* que expõe a média de sentimento por nota do usuário. O objetivo por trás da análise de gráficos do tipo *box plot* é verificar a distribuição dos dados. Assim, as conclusões possíveis ao analisar-se um *box plot* são: centro dos dados (a média ou mediana), a amplitude dos dados (máximo – mínimo), a simetria ou assimetria do conjunto de dados e a presença de *outliers*.

Figura 29 – Média de sentimento por avaliação segundo os léxicos para a EMPRESA 1

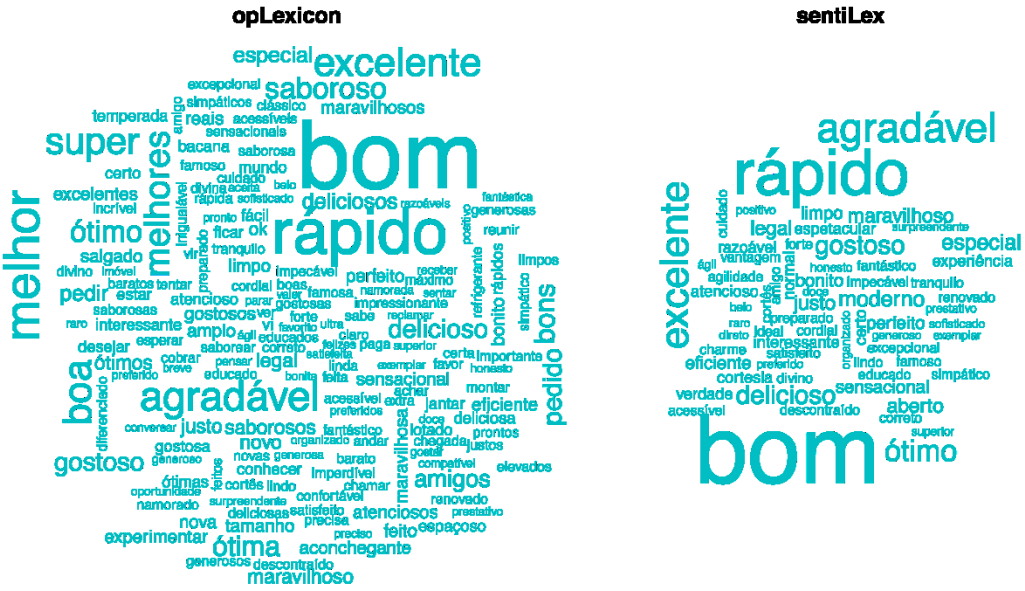


Fonte: Elaborado pelo autor.

No caso dos gráficos expostos há evidente diferença quando da aplicação de ambos os léxicos no *corpus* da EMPRESA 1, sugerindo-se uma melhor distribuição para o léxico opLexicon. Ambos apresentam certa assimetria, mas o desiquilíbrio é maior em relação ao léxico sentiLex. Outro fato curioso é que, no caso do léxico sentiLex, a presença de *outliers* é maior, se comparado ao opLexicon.

A próxima análise feita volta-se à presença de termos positivos classificados por cada léxico e apresentados lado a lado na Figura 30, no formato de nuvens de palavras.

Figura 30 – Nuvem de termos mais positivos por léxico para a EMPRESA 1

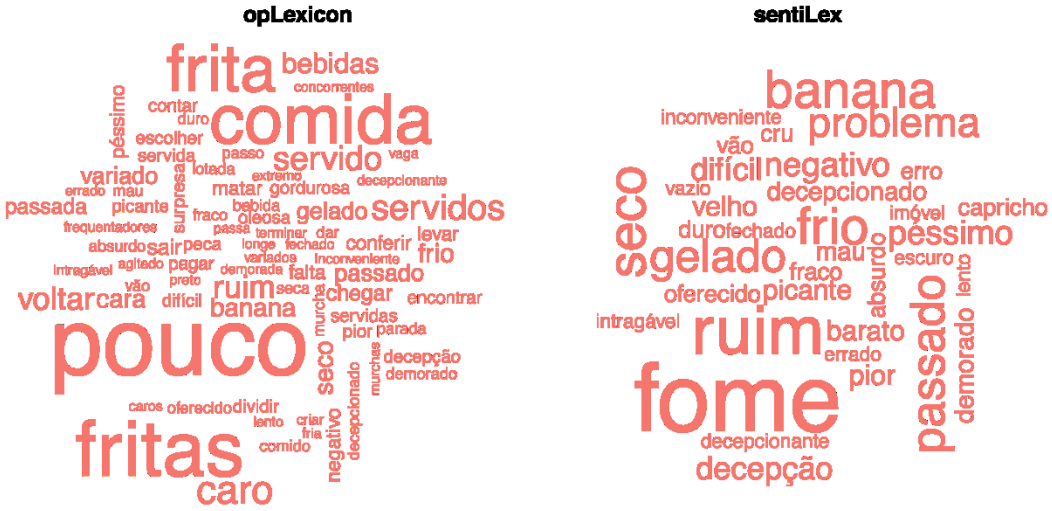


Fonte: Elaborado pelo autor.

Ao analisar-se ambas as nuvens, nota-se que há evidentes diferenças em relação aos termos classificados por ambos os léxicos. Na nuvem de termos positivos gerada pelo cruzamento com o léxico opLexicon, os termos ‘bom’, ‘agradável’ e ‘melhor’ destacam-se. Já na nuvem de termos positivos gerada pelo cruzamento com o léxico sentiLex, os termos que mais se destacam são ‘bom’, ‘rápido’, ‘agradável’ e ‘excelente’.

Da mesma forma, ao observar a nuvem de termos negativos gerada pelo cruzamento com o léxico opLexicon, os termos ‘pouco’, ‘comida’, ‘fritas’ e ‘frita’ se sobressaem. Quanto à nuvem negativa de termos oriunda do cruzamento com o léxico sentiLex, os termos ‘fome’ e ‘ruim’ destacam-se dos demais, seguidos por termos como ‘passado’, ‘seco’ e ‘frio’, entre outros, como verificado na Figura 31.

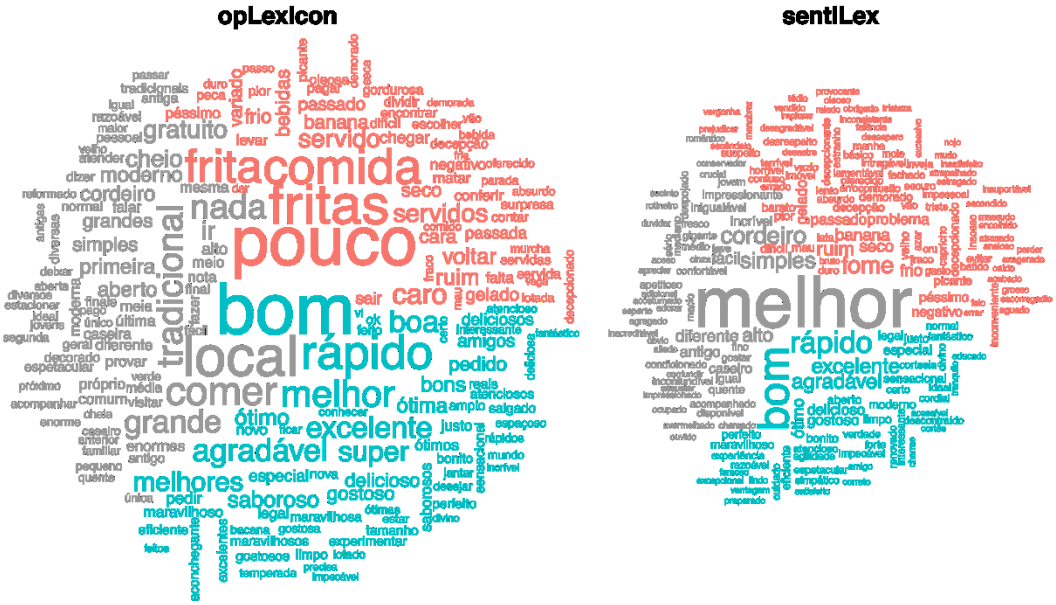
Figura 31 – Nuvem de termos negativos por léxico para a EMPRESA 1



Fonte: Elaborado pelo autor.

Outra forma de verificar a maneira como cada léxico classificou a polaridade dos termos presentes no *corpus* é a comparação da nuvem de termos, incluindo-se os termos positivos, negativos e neutros, conforme apresentada na Figura 32.

Figura 32 – Nuvem de termos positivos, negativos e neutros da EMPRESA 1



Fonte: Elaborado pelo autor.

Ao verificar a frequência de termos positivos, negativos e neutros, fica ainda mais evidente que o léxico pode impactar muito no resultado das análises finais. Considerando-se os termos presentes no léxico opLexicon verifica-se, como apontado anteriormente, que termos

importantes para o domínio estudado, como por exemplo ‘comida’, ‘servidos’ e ‘fritas’ são classificados como negativos, ao passo que termos como ‘grande’, ‘moderno’ e ‘gratuito’ são considerados neutros. Em relação ao léxico sentiLex, chama a atenção o fato de que o termo ‘melhor’ seja classificado como neutro e, considerando-se sua relevância e frequência, não parece uma classificação correta para este conjunto de dados.

Continuando as análises baseadas nas polaridades dos termos, realizou-se um cruzamento entre a relação dos termos positivos e negativos com a média das notas atribuídas pelos usuários, com o objetivo de descobrir quais termos mais positivos e mais negativos estão associados à média das avaliações atribuídas pelos usuários. O opLexicon foi o primeiro léxico a ser estudado por esta perspectiva, e os resultados são apresentados no Quadro 27.

Quadro 27 – Termos mais positivos e negativos com distribuições por documento e média de notas associadas segundo o léxico opLexicon para a EMPRESA 1

Termos mais positivos	Aparece em quantos documentos	Nota Média	Termos mais negativos	Aparece em quantos documentos	Nota Média
assíduos	2	5.0	estranho	2	1.0
atual	2	5.0	impessoal	2	1.0
certos	2	5.0	intragável	3	1.0
comidas	2	5.0	mediana	2	1.0
entregues	2	5.0	mole	2	1.0
famintos	2	5.0	terrível	2	1.0
fantástico	9	5.0	vazio	3	1.0
gastar	4	5.0	murchas	3	1.3
imbatível	2	5.0	horrível	2	1.5
imenso	2	5.0	péssimo	8	1.5

Fonte: Elaborado pelo autor

Da mesma forma, o resultado da classificação do léxico sentiLex é apresentado no Quadro 28, com destaque para os dez termos mais positivos e mais negativos.

Quadro 28 – Termos mais positivos e negativos com distribuições por documento e média de notas associadas segundo o léxico sentiLex para a EMPRESA 1

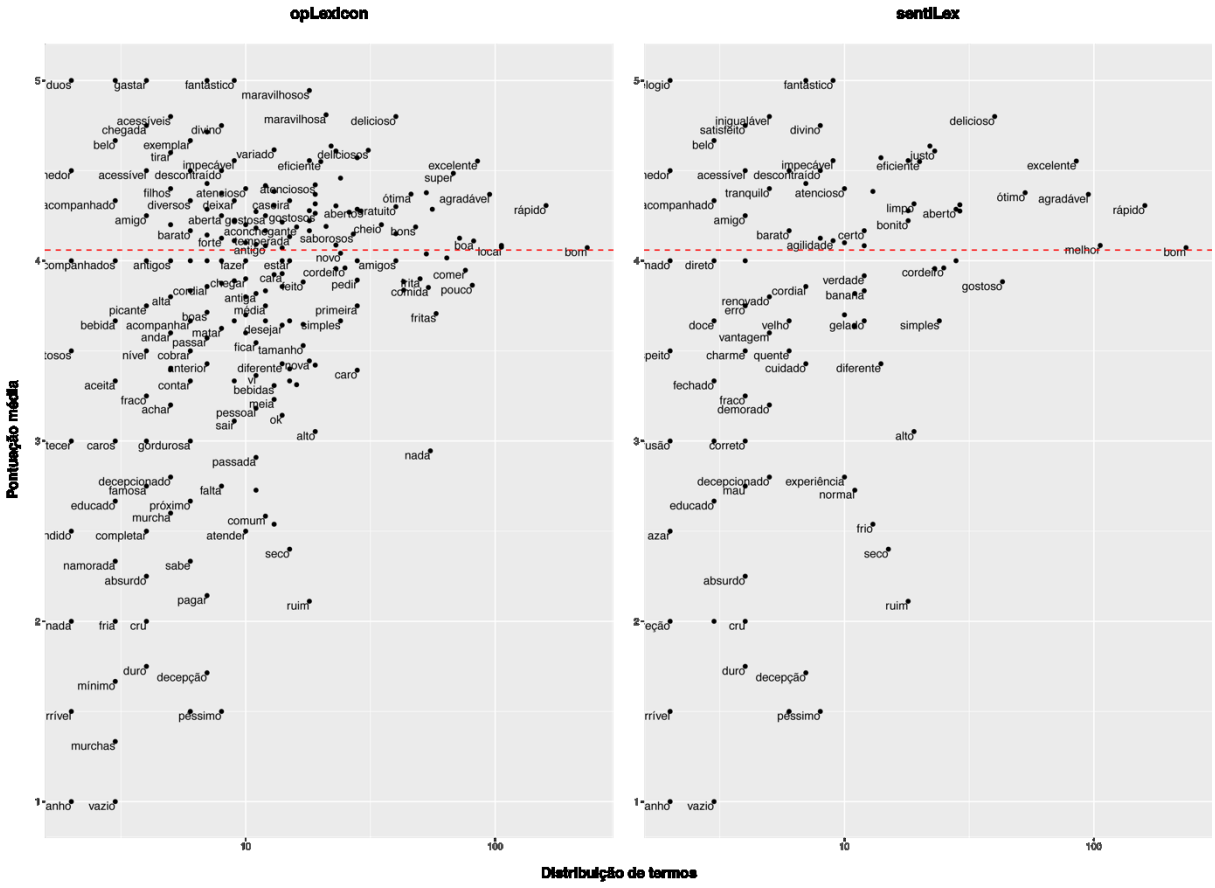
Termos mais positivos	Aparece em quantos documentos	Nota Média	Termos mais negativos	Aparece em quantos documentos	Nota Média
elogio	2	5.0	estranho	2	1.0
fantástico	9	5.0	impessoal	2	1.0
imbatível	2	5.0	intragável	3	1.0
impressionante	7	5.0	mole	2	1.0
inconfundível	2	5.0	terrível	2	1.0
jovem	2	5.0	vazio	3	1.0
macio	2	5.0	horrível	2	1.5
maravilha	2	5.0	péssimo	8	1.5
natural	2	5.0	pior	6	1.5
preferido	2	5.0	sério	2	1.5

Fonte: Elaborado pelo autor

Ao comparar ambos os resultados constantes nos Quadro 27 e Quadro 28, percebe-se que termos tanto negativos, quanto positivos constantes na seleção dos dez mais presentes variam pouco. Em outras palavras, por esta análise, ambos os léxicos desempenharam de forma semelhante. Mas é importante lembrar que esta análise é feita considerando-se a média da nota e a frequência de termos por documentos.

Ao considerar uma avaliação que pondere a média em relação à distribuição dos termos, talvez seja possível visualizar melhor as diferenças entre as classificações feitas por ambos os léxicos, conforme exposto na Figura 33.

Figura 33 – Distribuição de termos por avaliação segundo os léxicos para a EMPRESA 1



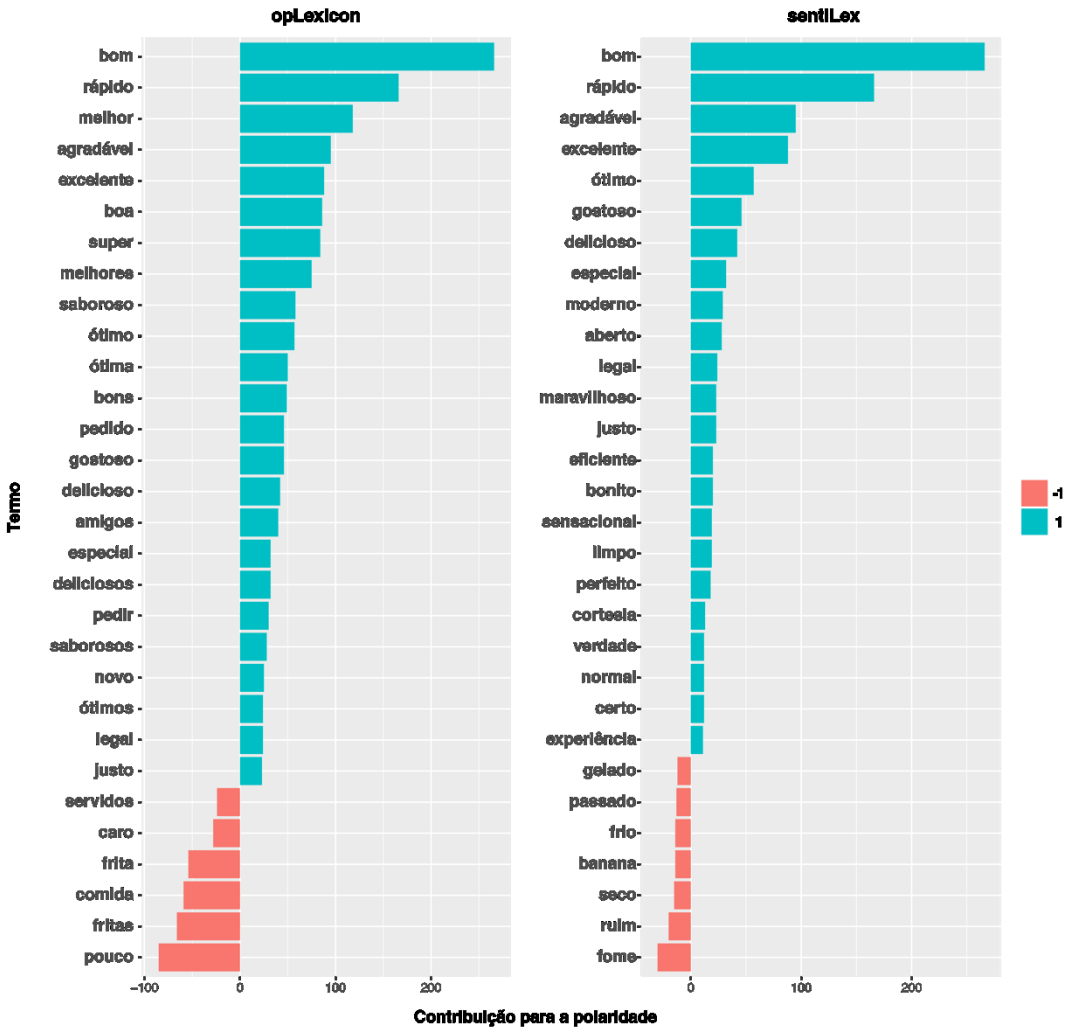
Fonte: Elaborado pelo autor.

Comparando-se os gráficos, pode-se verificar que os termos ‘rápido’, ‘excelente’, ‘agradável’ e ‘bom’ coincidem em ambos os léxicos e ocupam mais ou menos a mesma posição no gráfico. Isto deve-se à frequência de repetições e ao fato de estes termos serem alguns dos que constam de ambos os léxicos.

Mais uma vez a diferença entre ambos os léxicos aparece como um agravante nas análises. O gráfico gerado a partir dos termos constantes do opLexicon apresenta uma quantidade bem maior de termos classificados, o que pode significar análises mais ricas. Entretanto, é importante considerar mais uma vez, que ambos os léxicos não são específicos para o domínio abordado nesta pesquisa.

Em relação à diferença de classificações de ambos os léxicos, talvez a comparação fique mais clara ao proceder a comparação da frequência de termos positivos e negativos em oposição. Desta forma, espera-se que fique mais clara a contribuição dos termos para a Análise de Sentimentos, o que se apresenta na Figura 34.

Figura 34 – Termos positivos e negativos mais presentes no *corpus* da EMPRESA 1



Fonte: Elaborado pelo autor.

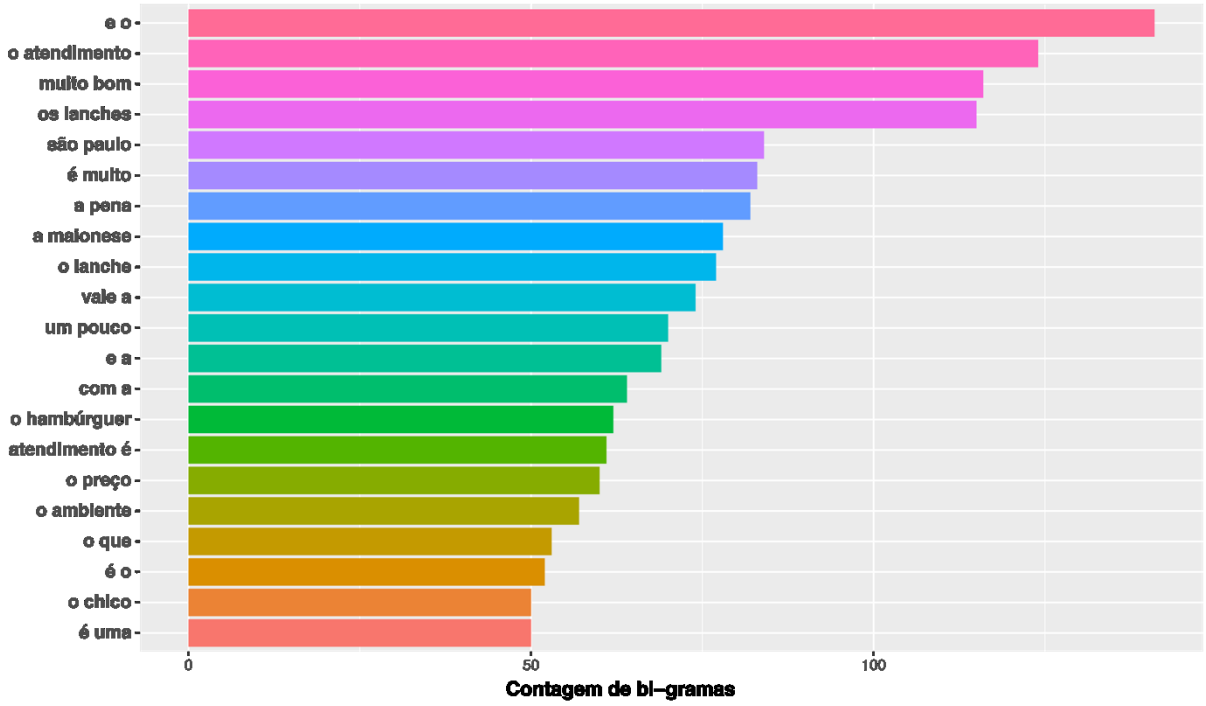
Quando colocados em perspectiva as frequências de termos positivos e negativos em oposição, fica ainda mais evidente que os léxicos avaliam termos de forma bem diferente. Um exemplo pode ser evidenciado pela comparação dos termos ‘melhor’, presente no léxico opLexicon e ‘agradável’, presente no léxico sentiLex. Ambos deveriam ser considerados positivos, mas não são considerados da mesma forma. Outro exemplo são os termos ‘caro’, do opLexicon e ‘banana’ do sentiLex, que são considerados de forma diferente. A informação de que algo é caro faz sentido como termo negativo para o opLexicon. Entretanto, o termo ‘banana’ não faz o menor sentido segundo a avaliação do léxico sentiLex, sobretudo se considerarmos que alguém está falando de um milk-shake de banana, por exemplo, o que é natural que aconteça, considerando-se o domínio desta pesquisa.

Até agora todas as análises levaram em consideração apenas uni-gramas, ou seja, um *token* formado por uma palavra. Entretanto, é possível obter análises mais ricas se olharmos

para termos que costumam aparecer juntos no *corpus*, ou seja, n-gramas. Esta pesquisa também considerou a utilização de bigramas para analisar a relação entre termos que aparecem juntos no documento. Os processos relacionados ao processamento dos bigramas já foram detalhados anteriormente no capítulo Procedimentos Metodológicos.

O gráfico apresentado na Figura 35 apresenta a frequência de termos que mais aparecem juntos em todo o *corpus*.

Figura 35 – Bigramas mais comuns em todo o *corpus* da EMPRESA 1

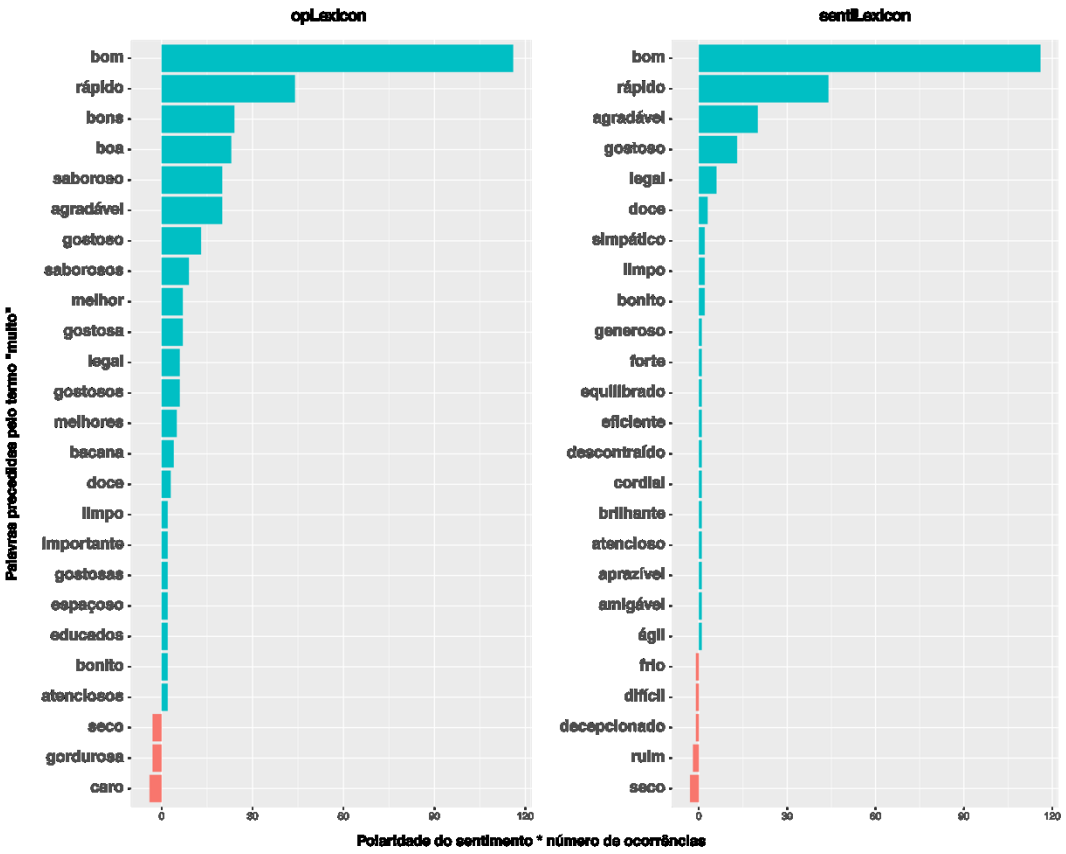


Fonte: Elaborado pelo autor.

Como explicado anteriormente, optou-se por não realizar a remoção de *stop words*, e o gráfico da Figura 35 mostra a presença de inúmeros termos considerados *stop words*. Conforme também exposto no capítulo que trata da metodologia, bi-gramas podem ser tratados como uni-gramas em ralação ao processo de Análise de Sentimentos.

Assim, após o cruzamento de ambos os termos dos bigramas com os léxicos, obtém-se como resultado termos que constam do léxico e que aparecem juntos, possibilitando descobrir, por exemplo, quais termos são precedidos por outros. A Figura 36 mostra os termos mais frequentes precedidos pela palavra 'muito', considerando-se ambos os léxicos empregados.

Figura 36 – Termos mais frequentes precedidos pela palavra 'muito' da EMPRESA 1



Fonte: Elaborado pelo autor.

Visando compreender as relações entre os termos que aparecem nas análises com bigramas, realizou-se um levantamento sobre quais destes termos aparecem juntos com mais frequência e em seguida, montou-se um grafo para apresentar os termos e a forma como eles se relacionam. Nesta análise nenhum dos léxicos foi usado, pois o único interesse é a relação entre os termos componentes e a forma como estes constam do *corpus*.

Para isso, procedeu-se a remoção das *stop words* de ambos os conjuntos de termos, resultando apenas em termos que aparecem juntos após a exclusão das *stop words*, o que pode ser visto no Quadro 29, que mostra as colunas com os termos e uma coluna contendo a frequência com que eles aparecem no *corpus*.

Quadro 29 –Dez bigramas mais frequentes no *corpus* sem *stop words* da EMPRESA 1

Termo1	Termo2	n
batata	frita	43
milk	shake	40
x	salada	32
atendimento	rápido	25
bom	atendimento	25
ambiente	agradável	23
cheese	salada	20
chico	hamburguer	20
batatas	fritas	19
chico	hambúrguer	18

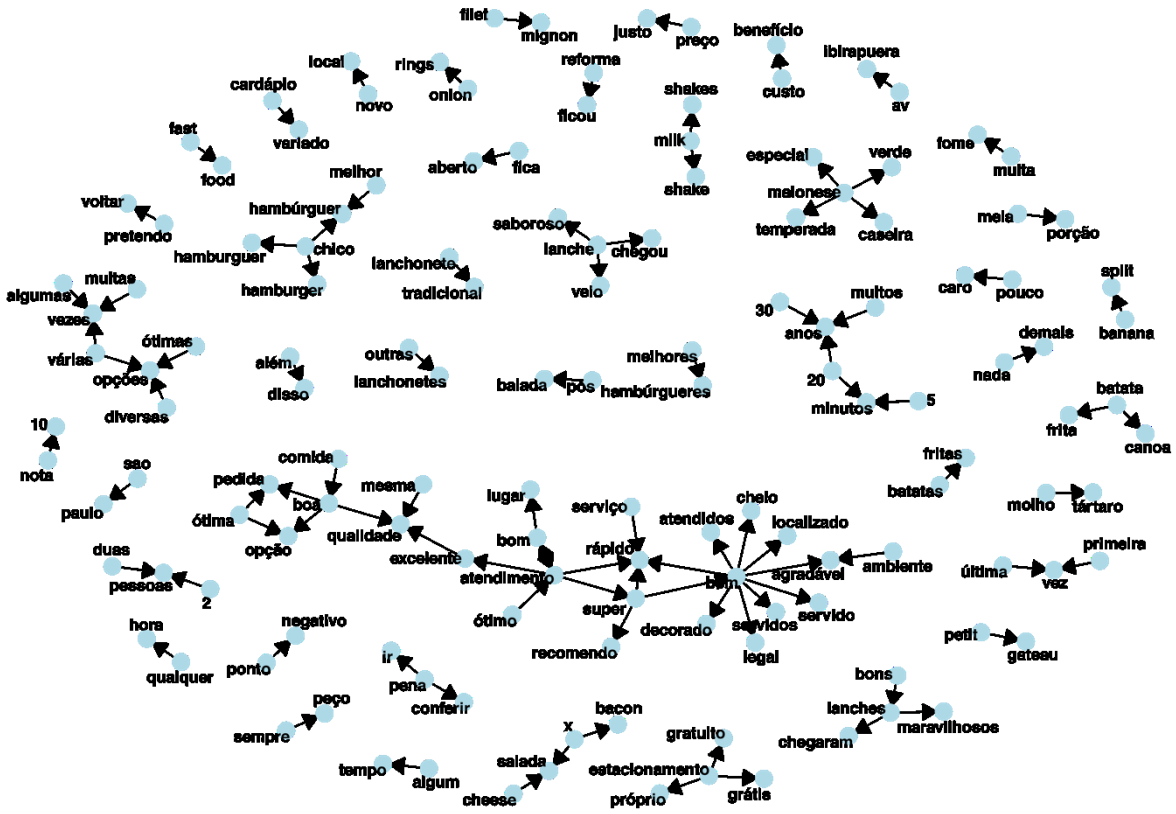
Fonte: Elaborado pelo autor

Vários termos ricos para análise começam a aparecer e, neste caso, referem-se principalmente à comida, ao lugar e ao atendimento, aspectos muito importantes para as análises. Uma curiosidade tem a ver como a maneira como algumas pessoas costumam escrever certos termos para referir-se à mesma coisa. Um exemplo disso são os termos ‘x’ e ‘cheese’, que em outro contexto significariam coisas diferentes, mas neste domínio significam exatamente a mesma coisa: queijo.

Nota-se que a seleção dos dez bigramas mais frequentes no *corpus* refere-se a aspectos positivos, o que, por sua vez, tem relação com a quantidade de opiniões e avaliações positivas, como já apresentado anteriormente.

O último passo desta parte da análise é representar a rede de palavras oriunda dos dados, conforme apresentado no Quadro 29. Para a geração da rede de palavras filtrou-se os bigramas mais comuns baseados no critério de frequência, ou seja, só consta do grafo os bigramas que se repetirem mais de cinco vezes (n > 5), o que se apresenta na Figura 37.

Figura 37 – Grafo de relações entre termos gerado a partir de bigramas da EMPRESA 1



Fonte: Elaborado pelo autor.

A rede de palavras mostra algumas coisas interessantes, como a relação entre diversos termos que até então não haviam sido expostos. Alguns exemplos disso são as relações entre termos como ‘estacionamento’, ‘gratuito’, ‘grátis’ e ‘próprio’; entre os termos ‘preço’ e ‘justo’ e ‘custo’ e ‘benefício’. Na verdade, a visualização de rede de palavras extraída a partir da relação entre bigramas tem o potencial de revelar diversos aspectos interessantes sobre o que as pessoas falam quando emitem suas opiniões.

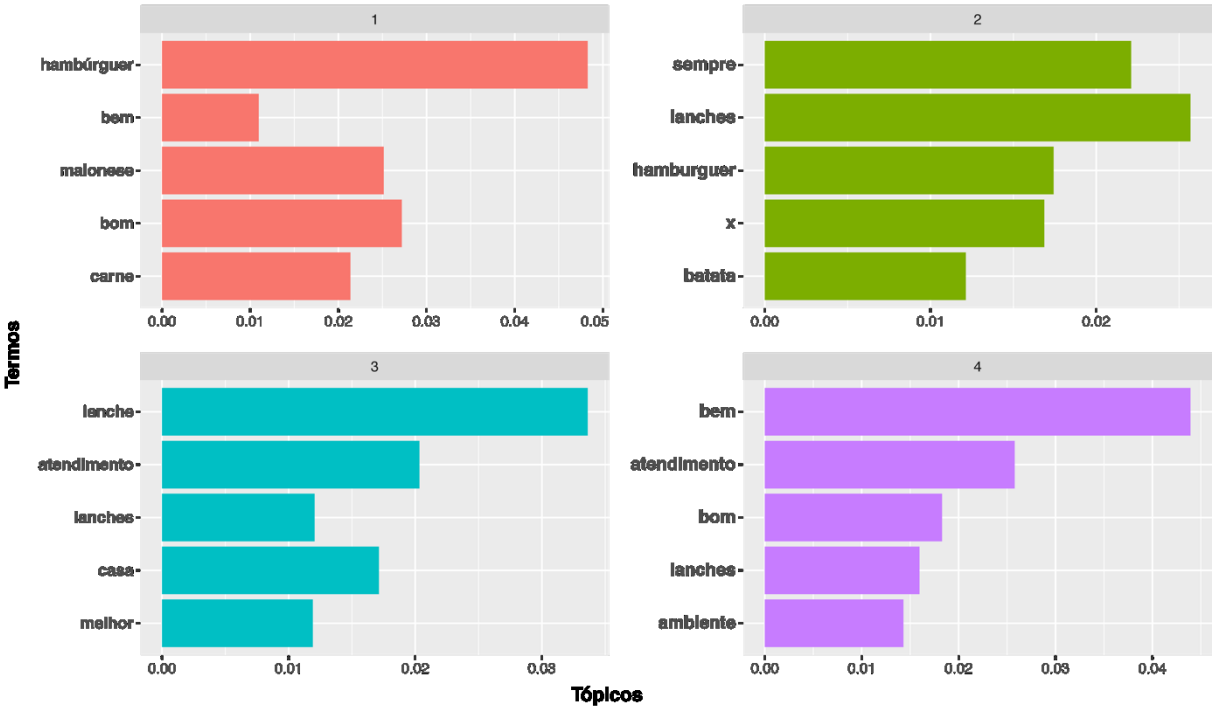
É importante ressaltar que esta rede é formada por termos que se repetem ao menos cinco vezes. Em outras palavras, são constantes em todo o conjunto de documentos. O que implica que muito mais pode ser extraído a partir da relação entre dois ou mais termos combinados.

Embora a imensa maioria dos comentários seja positiva, há sutilezas que podem e devem ser exploradas, como por exemplo a relação de termos aparentemente negativos, como ‘bem + cheio’, ‘pouco + caro’ e ‘última + vez’. A primeira relação pode referir-se ao fato do lugar estar sempre muito cheio, a segunda ao fato dos preços serem um pouco altos e a última, a uma experiência frustrante que não deverá gerar o retorno do cliente ao estabelecimento.

Não obstante, pode-se supor que, considerando a distribuição de opiniões positivas e negativas, há muito pouco o que destacar como pontos negativos no conjunto de dados analisado.

Para finalizar a análise dos dados da EMPRESA 1, aplicou-se a técnica de Modelagem de Tópicos, tendo como parâmetro a geração de quatro tópicos com suas devidas distribuições de termos, o que pode ser visto na Figura 38.

Figura 38 – Resultado da Modelagem de Tópicos da EMPRESA 1



Fonte: Elaborado pelo autor.

Como a Modelagem de Tópicos não leva em consideração nenhum dos léxicos e avalia o *corpus* inteiro, a única operação realizada foi a remoção de *stop words*. Dada a distribuição encontrada pelo modelo, aparentemente o Tópico 1 fala sobre ‘comida’, o que se evidencia pela presença de termos como ‘hambúrguer’, ‘maionese’ e ‘carne’. O Tópico 2 também fala sobre comida, dada a presença de termos como ‘lanches’, ‘hambúrguer’, ‘x’ e ‘batata’. Entretanto os Tópicos 3 e 4 falam sobre uma mistura de assuntos que pode ser compreendida como parte relacionado à comida, parte relacionado ao atendimento e ao local.

Conclui-se, portanto, que os clientes da EMPRESA 1 falam muito bem do restaurante. Quando comentam sobre o restaurante, falam muito sobre a comida e em seguida, sobre o local e sobre o atendimento.

4.2 Análise dos dados da EMPRESA 2

A EMPRESA 2 é um restaurante fundado em 1965 e tem como principais produtos sanduíches e pratos rápidos. Além do TripAdvisor, a EMPRESA 2 possui apenas o Facebook como rede social ativa, no qual possui 24.068 seguidores. O restaurante está no TripAdvisor desde novembro de 2010.

Seguindo a estrutura já apresentada na Empresa 1, o processo de análise começa com o pré-processamento dos dados, cujos resultados podem ser vistos no Quadro 30.

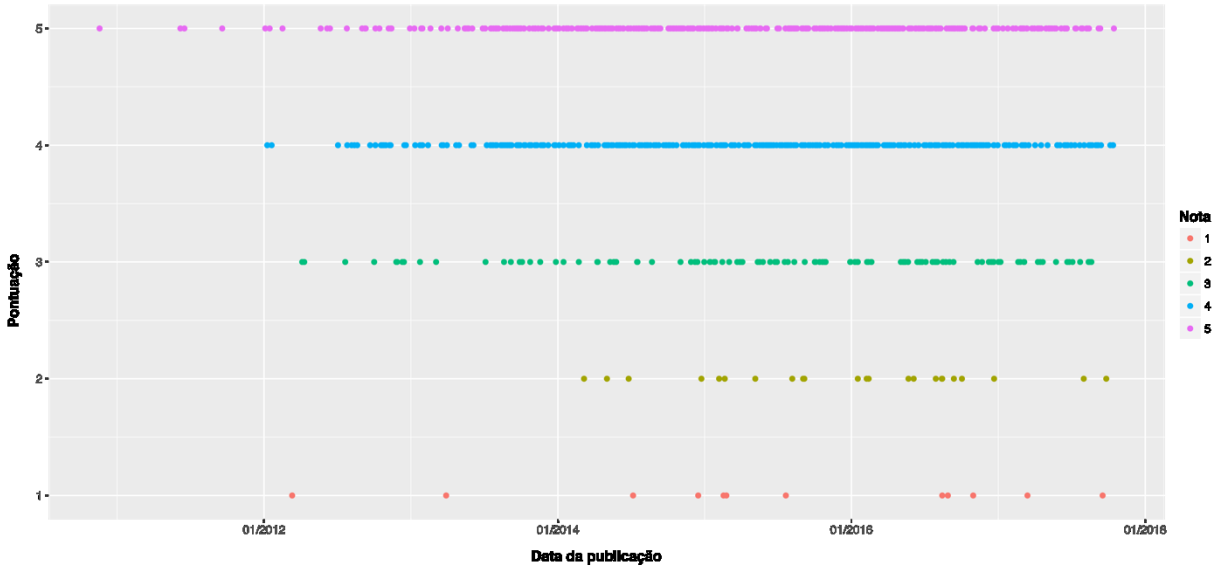
Quadro 30 – Sumarização dos dados da EMPRESA 2 após fase de pré-processamento

Procedimento	Quantidade de itens após procedimento
Entrada	953 documentos
Tokenização	39.951 palavras
Remoção de *stop words*	22.950 palavras
Remoção de números e caracteres especiais	22.458 palavras

Fonte: Elaborado pelo autor

Como resultado das etapas do pré-processamento, obteve-se um produto de 22.458 palavras sem *stop words*, distribuídas nos 953 documentos da EMPRESA2. A análise dos dados começa com a visualização da distribuição das notas ao longo do tempo, apresentada na Figura 39.

Figura 39 – Distribuição das notas ao longo do tempo para a EMPRESA 2

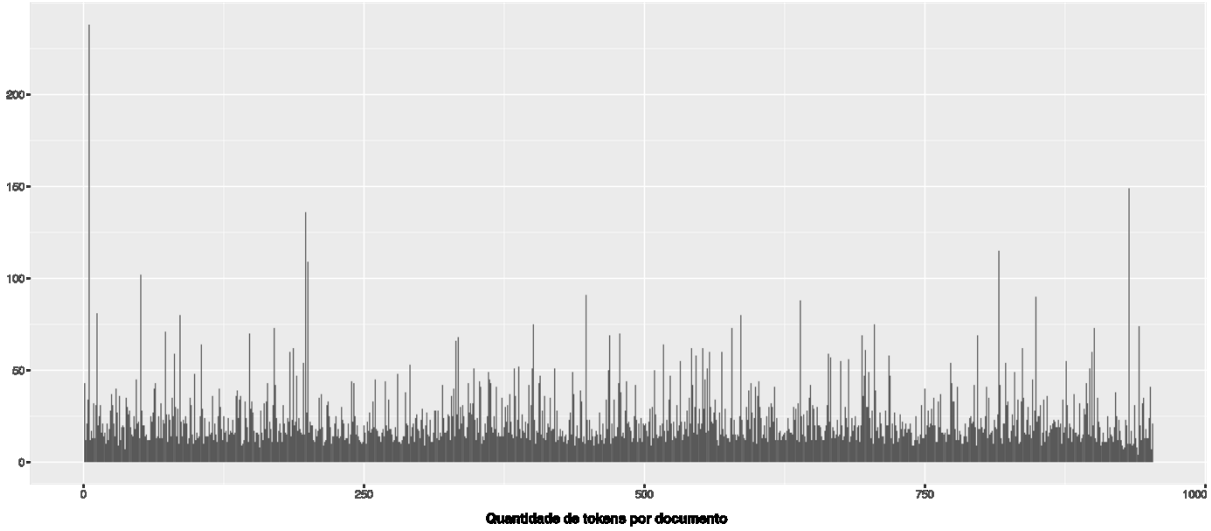


Fonte: Elaborado pelo autor.

Este é outro caso em que a quantidade de notas positivas (4 e 5) supera as demais notas atribuídas, o que denota a preferência das pessoas pelo estabelecimento. Mesmo assim, da mesma forma que realizado com os dados da Empresa 1, aprofundar-se-ão as análises visando obter uma melhor visão sobre o que os clientes falam quando fazem comentários positivos ou negativos ao estabelecimento EMPRESA 2.

A próxima análise representa o tamanho médio das avaliações por meio de um histograma que reflete a quantidade de termos por documento, conforme mostra a Figura 40.

Figura 40 – Quantidade de termos por documento da EMPRESA 2



Fonte: Elaborado pelo autor.

Analisando-se as repetições de palavras, nota-se que, pela repetição absoluta dos termos no *corpus*, os clientes falam muito sobre o atendimento (374 repetições), sobre a comida (mais de 780 repetições, considerando a soma dos termos maionese, lanches e lanche) e sobre o local (mais de 340 repetições).

Quadro 31 – Dez termos mais presentes em todo o *corpus* da EMPRESA 2

Termo	Quantidade de repetições no *corpus*
atendimento	374
bem	326
bom	317
maionese	317
lanches	275
sempre	247
lanche	195
estacionamento	178
melhor	167
ambiente	163

Fonte: Elaborado pelo autor

O Quadro 32 considera os dez termos com mais repetições por documento, a fim de proporcionar uma visão diferente sobre a frequência de alguns termos.

Quadro 32 – Dez termos com mais repetições por documento da EMPRESA 2

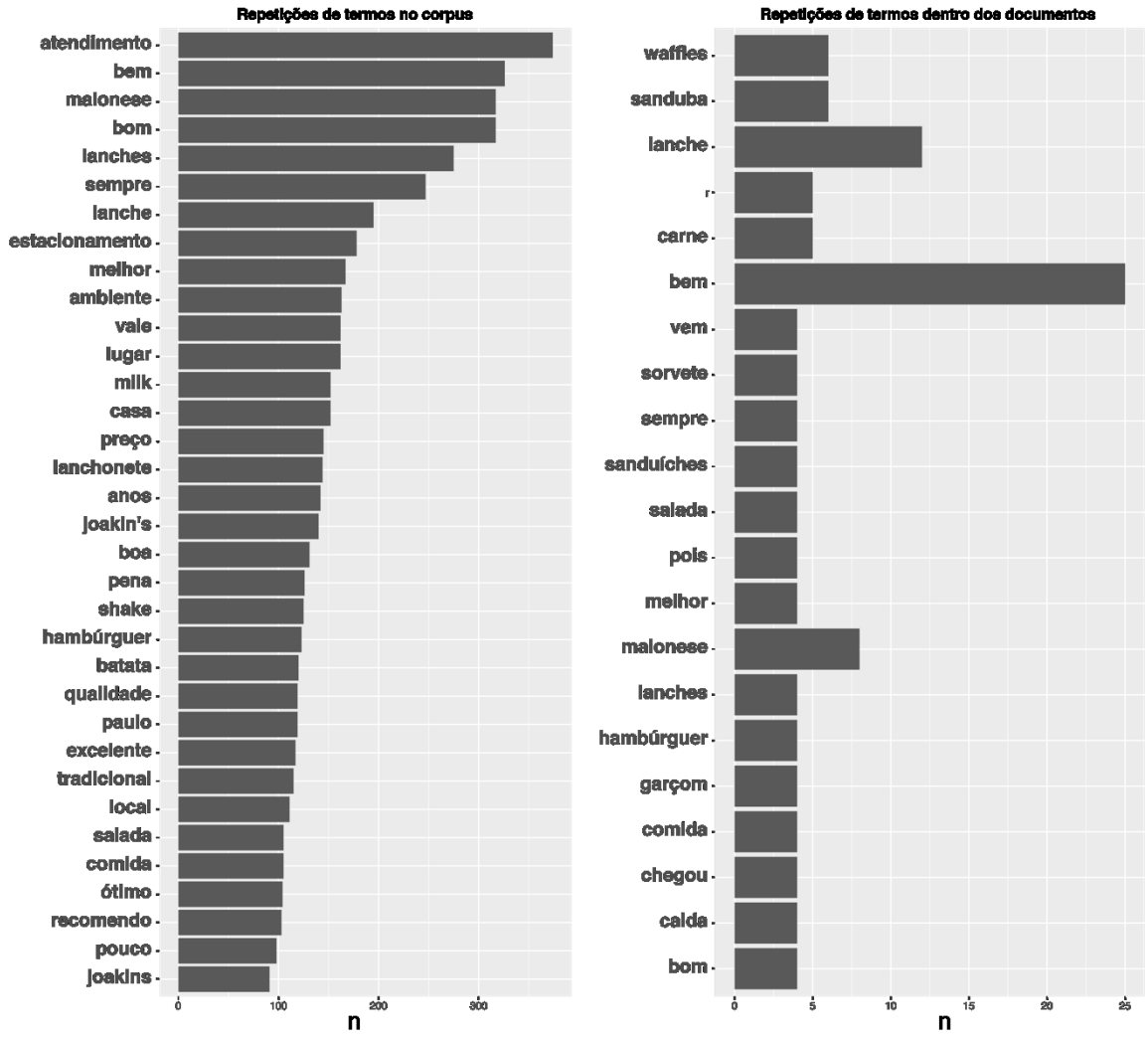
Termo	Quantidade de repetições por documento
sanduba	6
lanche	6
lanche	6
waffles	6
r	5
carne	5
bem	5
bem	4
vem	4
pois	4

Fonte: Elaborado pelo autor

Nota-se a repetição de termo ‘sanduba’ e a presença de um termo alheio à análise (a letra ‘r’), que muito provavelmente refere-se ao termo ‘rua’, indicando a localização do restaurante, e passou intacto pela remoção de *stop words*.

O gráfico apresentado na Figura 41 apresenta a plotagem de repetição de termos no *corpus* e dentro dos documentos. Nota-se que, novamente, o termo ‘bem’ é uma constante em todas as análises. Entretanto, os motivos para sua manutenção já foram explicados anteriormente.

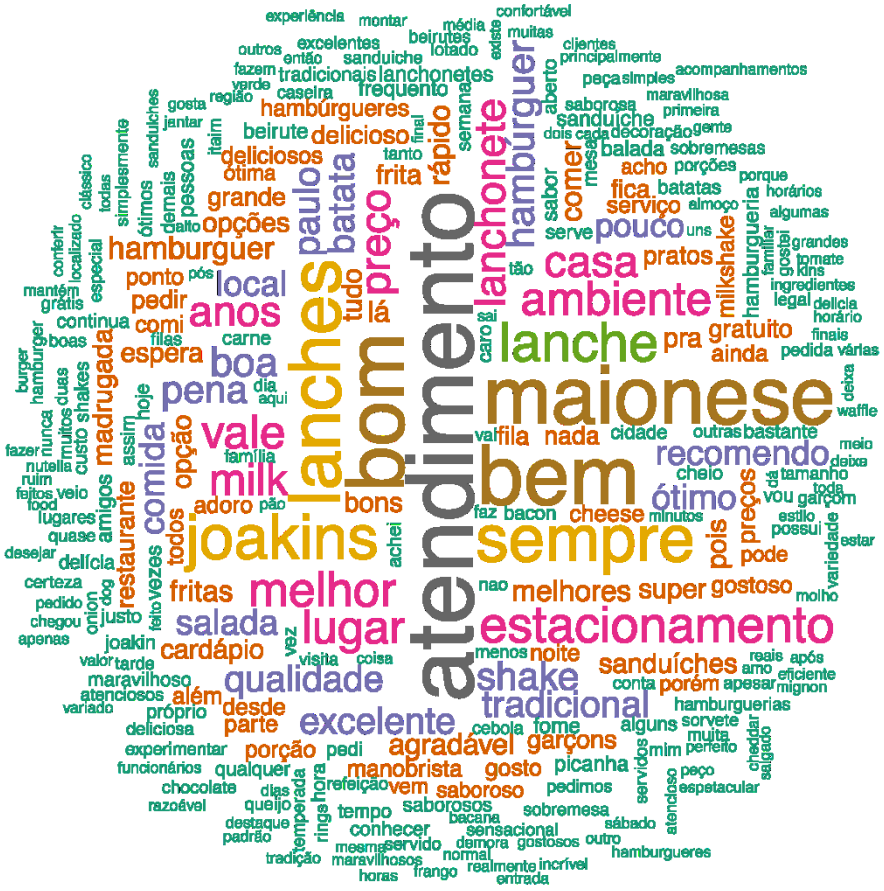
Figura 41 – Repetição de termos no *corpus* e dentro dos documentos da EMPRESA 2



Fonte: Elaborado pelo autor.

Outra forma de avaliar a presença de determinados termos em relação ao quanto estes termos se repetem num *corpus* é contrastar termos mais frequentes com termos menos frequentes, aqui apresentados como uma nuvem de palavras, cujo resultado é exposto na Figura 42.

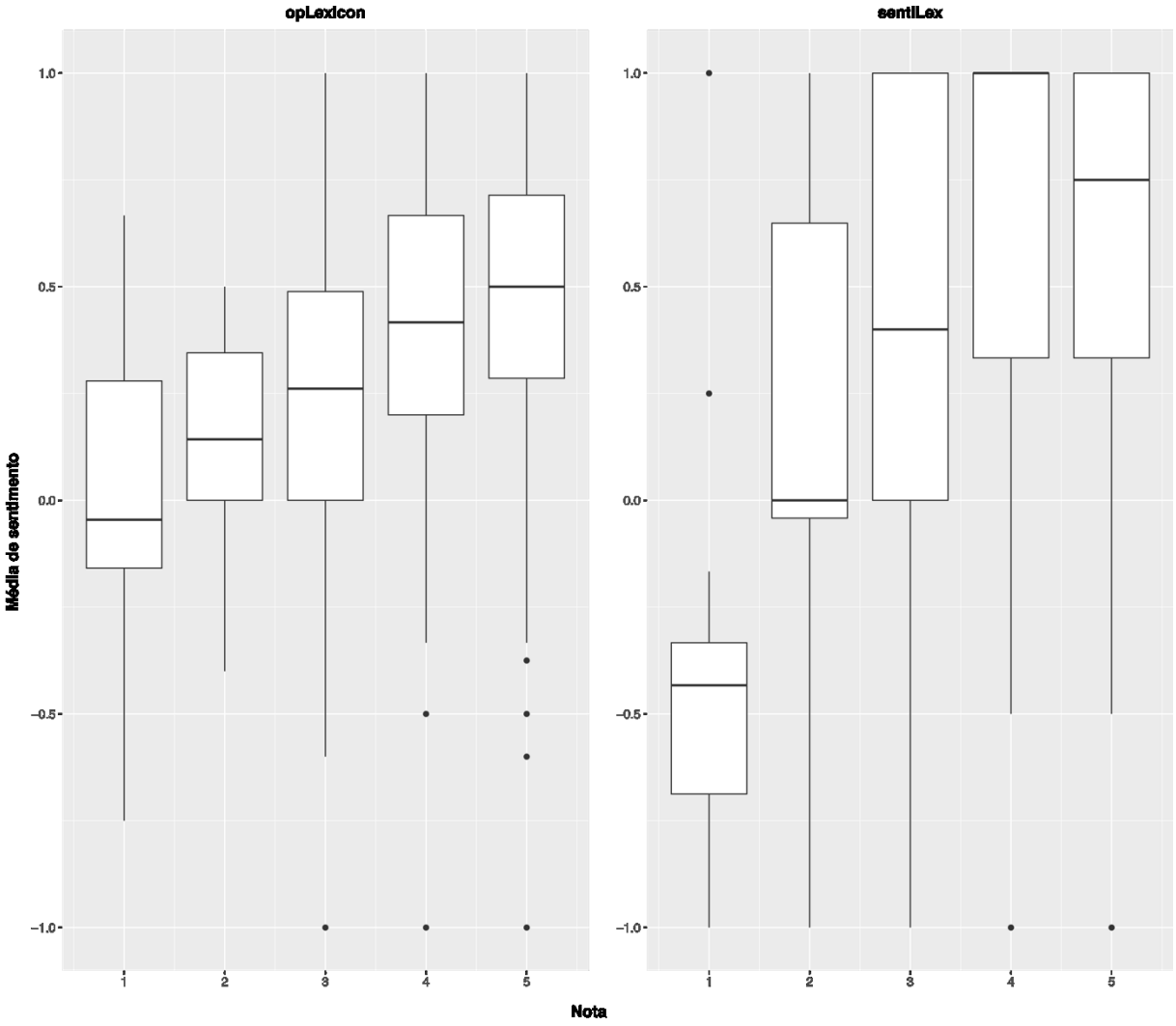
Figura 42 – Nuvem de termos mais frequentes da EMPRESA 2



Fonte: Elaborado pelo autor.

A nuvem de palavras apresentada na Figura 42 corrobora o que foi encontrado em relação à frequência de termos. Procede-se agora à Análise de Sentimentos com o objetivo de avaliar as inclinações das postagens dos usuários. A primeira análise cruza os dados de pontuação atribuída pelo usuário numa escala de 1 a 5 e a classificação de sentimentos atribuída pelo léxico. A Figura 43 mostra um gráfico do tipo *box plot* que consiste na média de sentimento por nota do usuário.

Figura 43 – Média de sentimento por avaliação segundo os léxicos para a EMPRESA 2

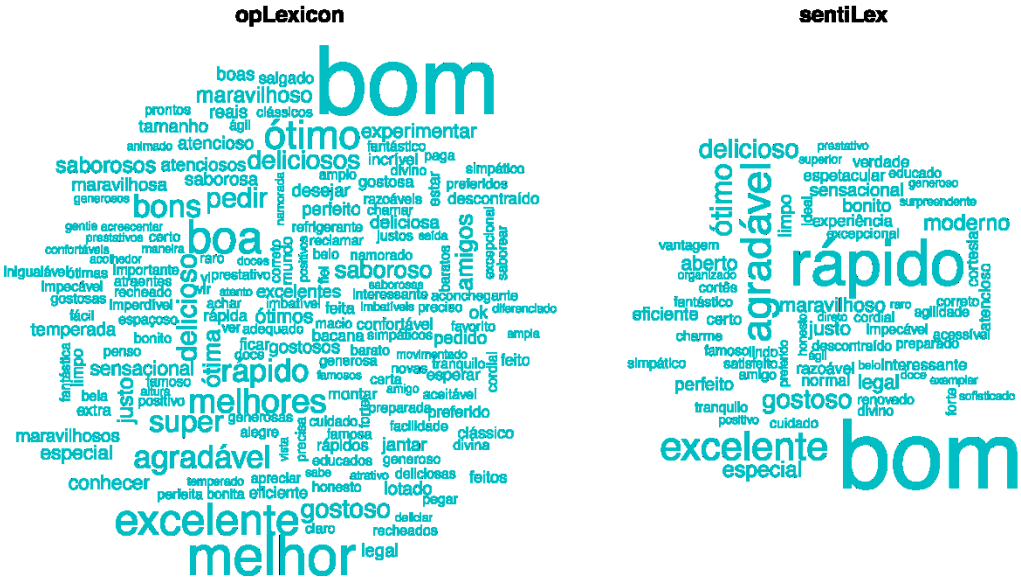


Fonte: Elaborado pelo autor.

Novamente constata-se a diferença proveniente da aplicação de ambos os léxicos no *corpus* da EMPRESA 2, com melhor distribuição para o léxico opLexicon. Ambos apresentam certa assimetria, mas o desiquilíbrio é maior em relação ao léxico sentiLex. Nota-se que neste caso, a presença de *outliers* é maior no léxico opLexicon.

A próxima análise se dá em relação à presença de termos positivos classificados por cada léxico e apresentados lado a lado na Figura 44.

Figura 44 – Nuvem de termos mais positivos por léxico para a EMPRESA 2

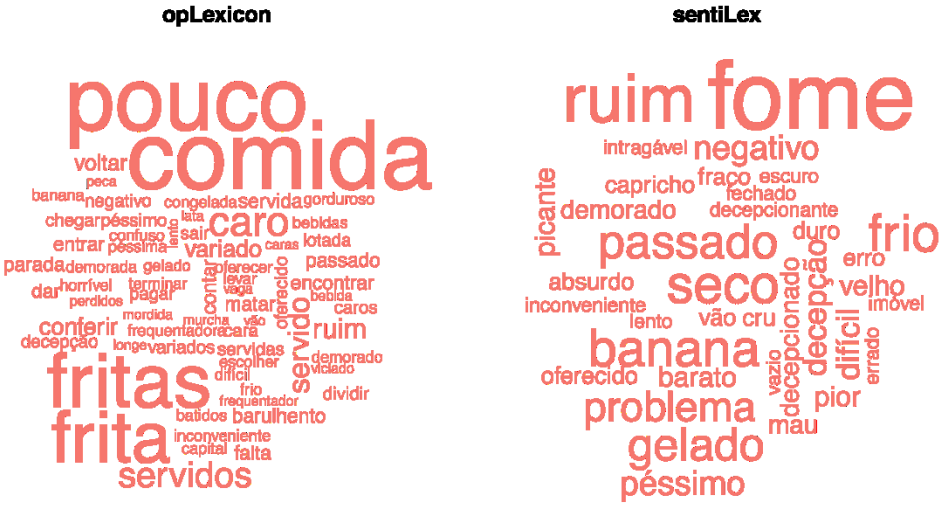


Fonte: Elaborado pelo autor.

Na nuvem de termos positivos gerada pelo cruzamento com o léxico opLexicon, os termos ‘bom’, ‘melhor’, ‘rápido’ e ‘ótimo’ destacam-se. Já na nuvem de termos positivos gerada pelo cruzamento com o léxico sentiLex, os termos que mais se destacam são ‘bom’, ‘ótimo, ‘agradável’, ‘excelente’ e ‘rápido’.

A Figura 45 apresenta a nuvem de termos negativos gerada pelo cruzamento com o léxico opLexicon, onde destacam-se os termos ‘comida’, ‘fritas/frita’, ‘pouco’ e ‘caro’. Quanto à nuvem negativa de termos oriunda do cruzamento com o léxico sentiLex, os termos ‘fome’ e ‘ruim’ destacam-se dos demais, seguidos por termos como ‘barato’, ‘passado’, ‘barulhento’ e ‘decepção’, entre outros, como verificado na Figura 45.

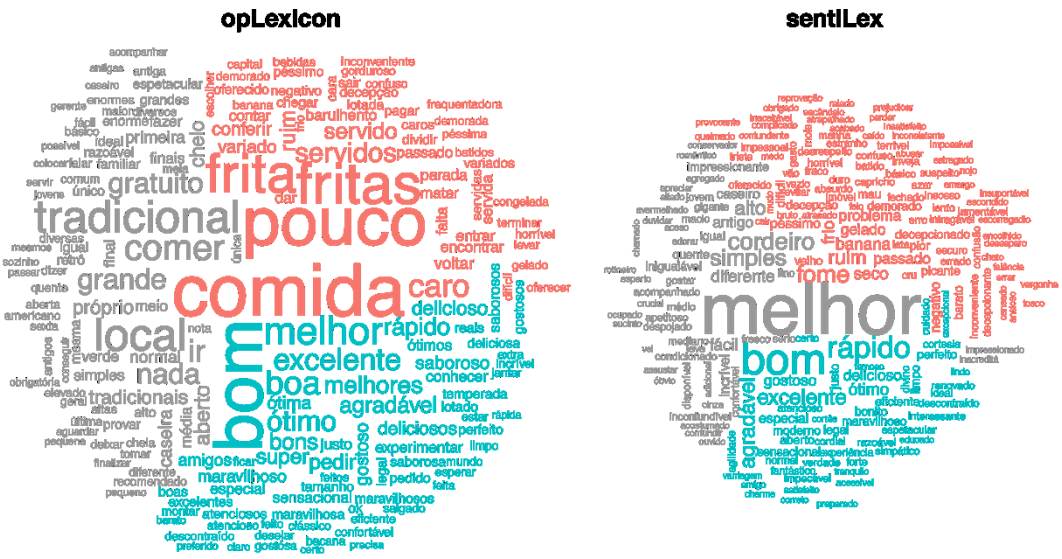
Figura 45 – Nuvem de termos negativos por léxico para a EMPRESA 2



Fonte: Elaborado pelo autor.

A Figura 46 apresenta a distribuição de termos com polaridade presentes no *corpus,* incluindo-se termos positivos, negativos e neutros.

Figura 46 – Nuvem de termos positivos, negativos e neutros da EMPRESA 2



Fonte: Elaborado pelo autor.

Considerando-se os termos presentes no léxico opLexicon, verifica-se que termos importantes para o domínio estudado, como por exemplo, ‘lugar’, ‘comer’ e ‘tradicional’ são considerados neutros, embora sejam significativos, considerando-se o histórico do restaurante.

Em relação ao léxico sentiLex, chama a atenção que o termo ‘melhor’ seja classificado como neutro e mais uma vez se repita tanto em mais um conjunto de dados.

A próxima análise envolve um cruzamento entre a relação dos termos positivos e negativos com a média das notas atribuídas pelos usuários, buscando-se descobrir quais termos mais positivos e mais negativos estão associados com a média das avaliações conferidas pelos usuários. O opLexicon foi o primeiro léxico a ser estudado por esta perspectiva, e os resultados por ser verificados no Quadro 33.

Quadro 33 – Termos mais positivos e negativos com distribuições por documento e média de notas associadas segundo o léxico opLexicon para a EMPRESA 2

Termos mais positivos	Aparece em quantos documentos	Nota Média	Termos mais negativos	Aparece em quantos documentos	Nota Média
altamente	2	5.0	errada	2	1.0
amantes	2	5.0	horrível	2	1.25
apaixonada	3	5.0	péssima	3	1.25
assíduo	2	5.0	decepção	2	1.5
atuais	2	5.0	montado	2	1.5
breve	2	5.0	ruins	2	1.5
comemorar	2	5.0	péssimo	3	1.6
comentar	2	5.0	comprado	3	2.0
comparados	2	5.0	comuns	2	2.0
deliciar	3	5.0	cortar	8	2.0

Fonte: Elaborado pelo autor

Da mesma forma, o resultado da classificação do léxico sentiLex é apresentado no Quadro 34, com destaque para os dez termos mais positivos e mais negativos.

Quadro 34 – Termos mais positivos e negativos com distribuições por documento e média de notas associadas segundo o léxico sentiLex para a EMPRESA 2

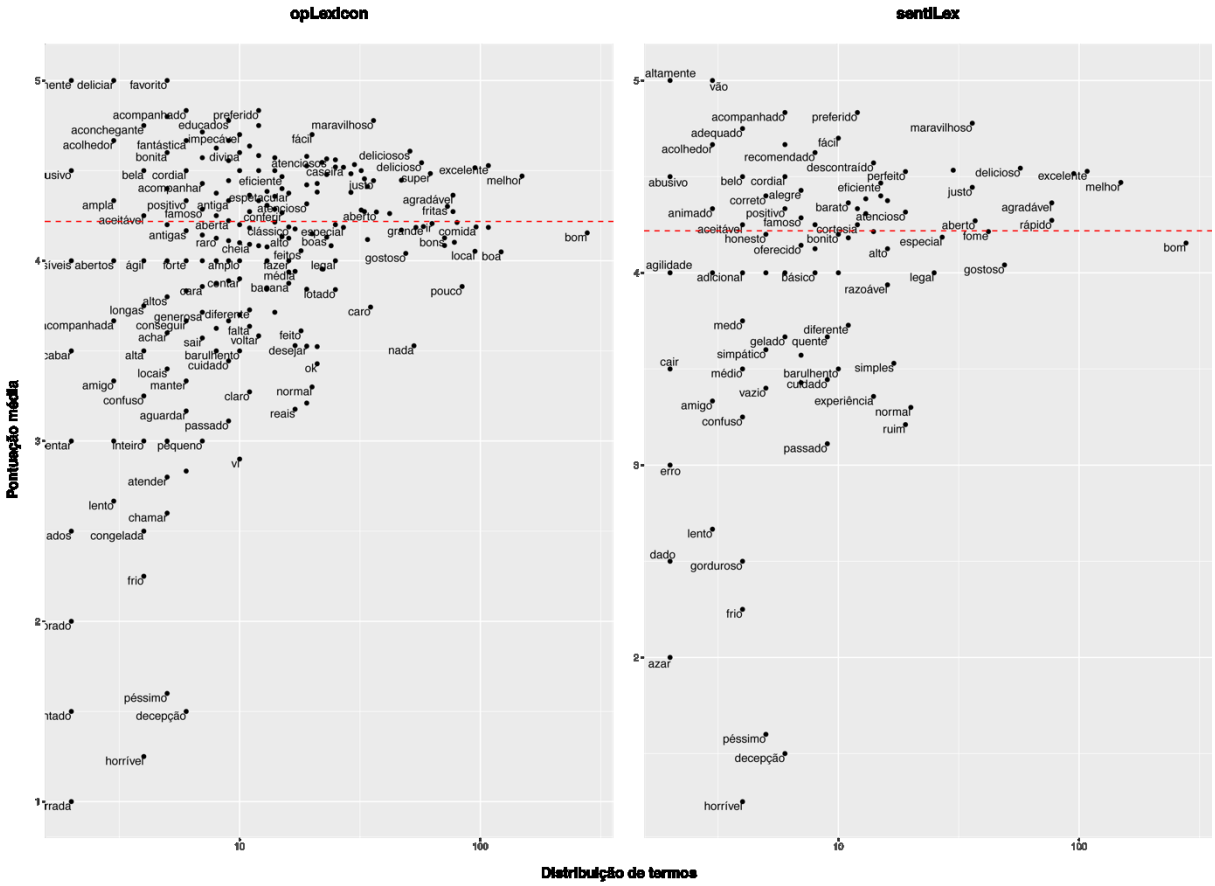
Termos mais positivos	Aparece em quantos documentos	Nota Média	Termos mais negativos	Aparece em quantos documentos	Nota Média
altamente	2	5.0	horrível	4	1.2
assíduo	2	5.0	decepção	6	1.5
errar	2	5.0	péssimo	5	1.6
flexibilidade	2	5.0	azar	2	2.0
inconfundível	2	5.0	comprado	2	2.0
ligeiro	2	5.0	cru	2	2.0
responsável	2	5.0	decepcionado	2	2.0
vão	3	5.0	decepcionante	2	2.0
verdadeiro	2	5.0	estranho	2	2.0
acompanhado	6	4.8	pior	2	2.0

Fonte: Elaborado pelo autor

Ao comparar-se os resultados do Quadro 33 e Quadro 34 observa-se que os termos, tanto negativos quanto positivos, constantes na seleção dos dez mais presentes variam bastante, com destaque para os termos ‘decepcionado’, ‘decepcionante’ e ‘cru’ que aparecem na análise do sentiLex, mas não no opLexicon, e parecem fazer sentido para descrever situações importantes ao contexto do *corpus* analisado.

Avaliando-se a distribuição dos termos em relação à pontuação média, é possível visualizar melhor as diferenças entre as classificações feitas por ambos os léxicos, conforme mostra a Figura 47.

Figura 47 – Distribuição de termos por avaliação segundo os léxicos para a EMPRESA 2

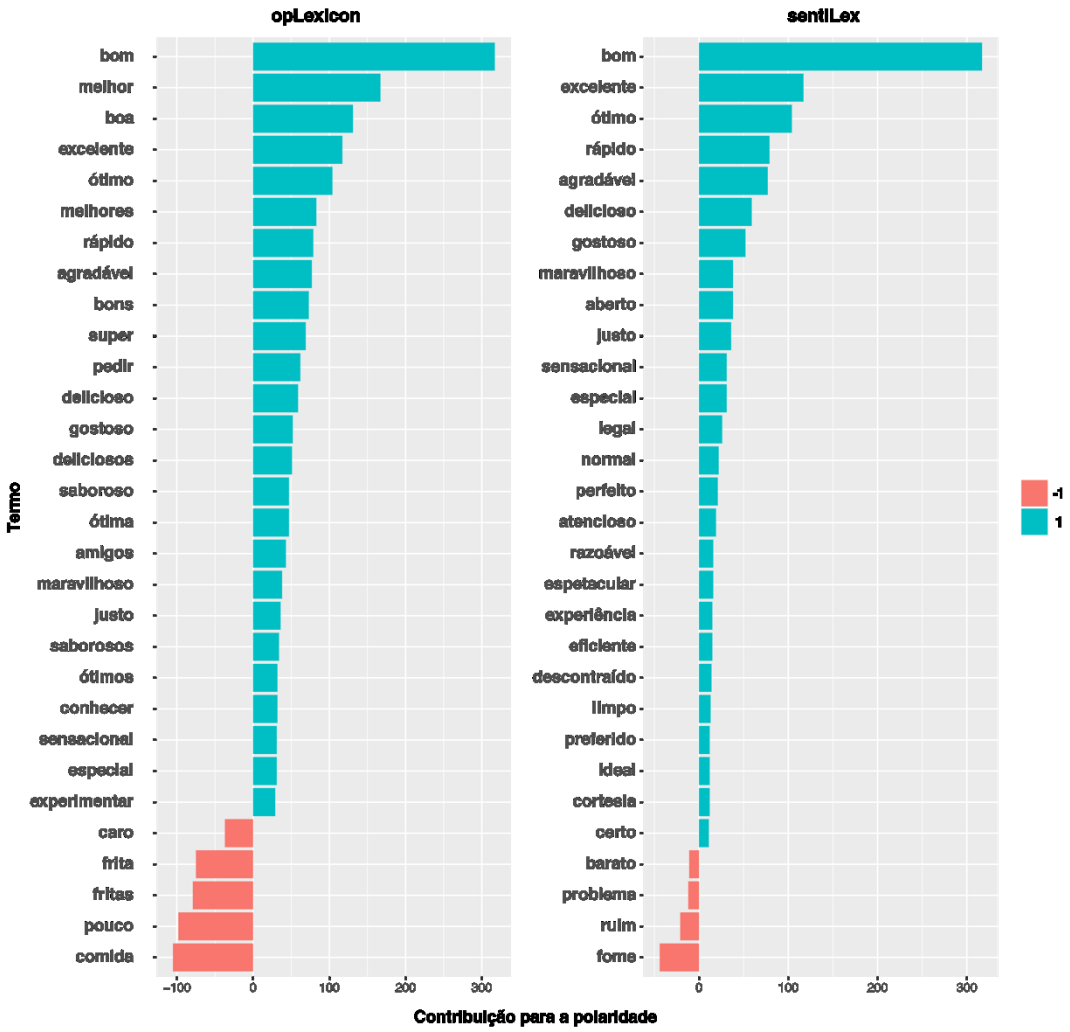


Fonte: Elaborado pelo autor.

Ao comparar os gráficos, nota-se novamente a presença de termos comuns a ambos os léxicos, como ‘agradável’, ‘excelente’ e ‘bom’. Entretanto, o opLexicon parece descrever melhor o que se passa com o conjunto de dados.

Para fins de comparação, a Figura 48 mostra a frequência de termos positivos e negativos classificados por ambos os léxicos em oposição.

Figura 48 – Termos positivos e negativos mais presentes no *corpus* da EMPRESA 2

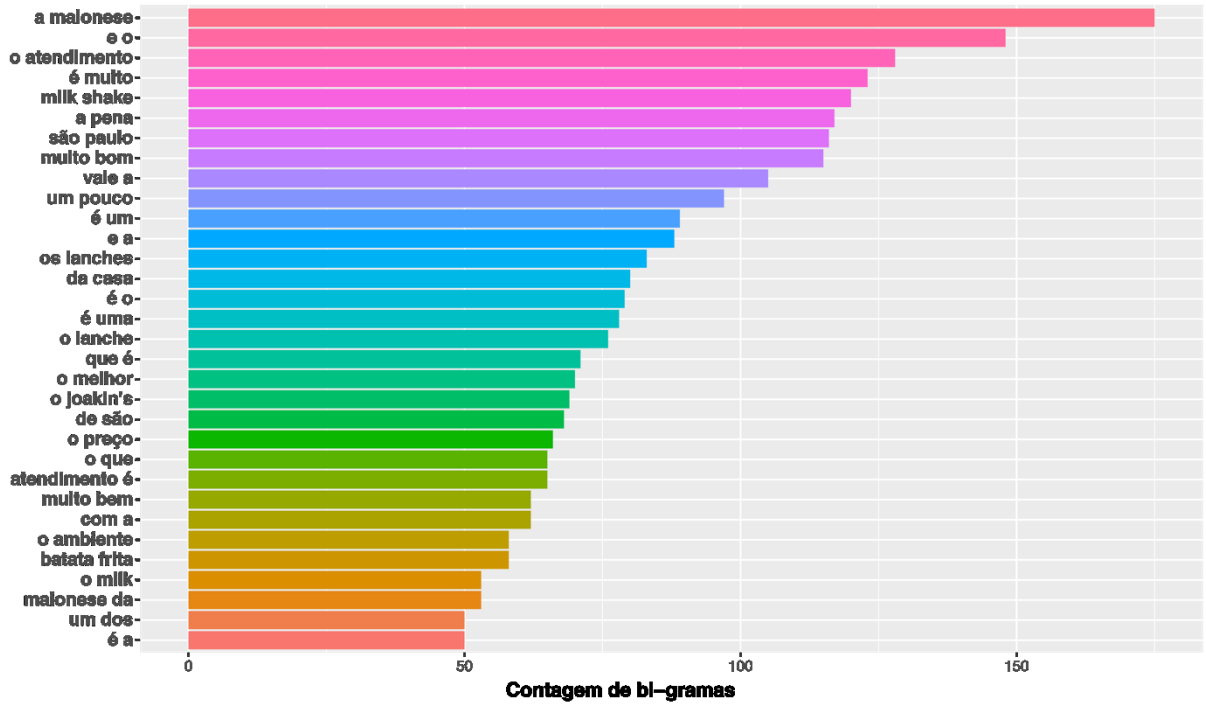


Fonte: Elaborado pelo autor.

O gráfico da Figura 48 mostra que vários termos coincidem, como os positivos ‘bom’, ‘rápido’ e ‘agradável’, mas alguns negativos merecem atenção, como os termos ‘caro’ e ‘problema’, que se repetem algumas vezes.

As próximas análises consideram bigramas, começando pela plotagem de termos que mais aparecem juntos em todo o *corpus*, como mostra a Figura 49.

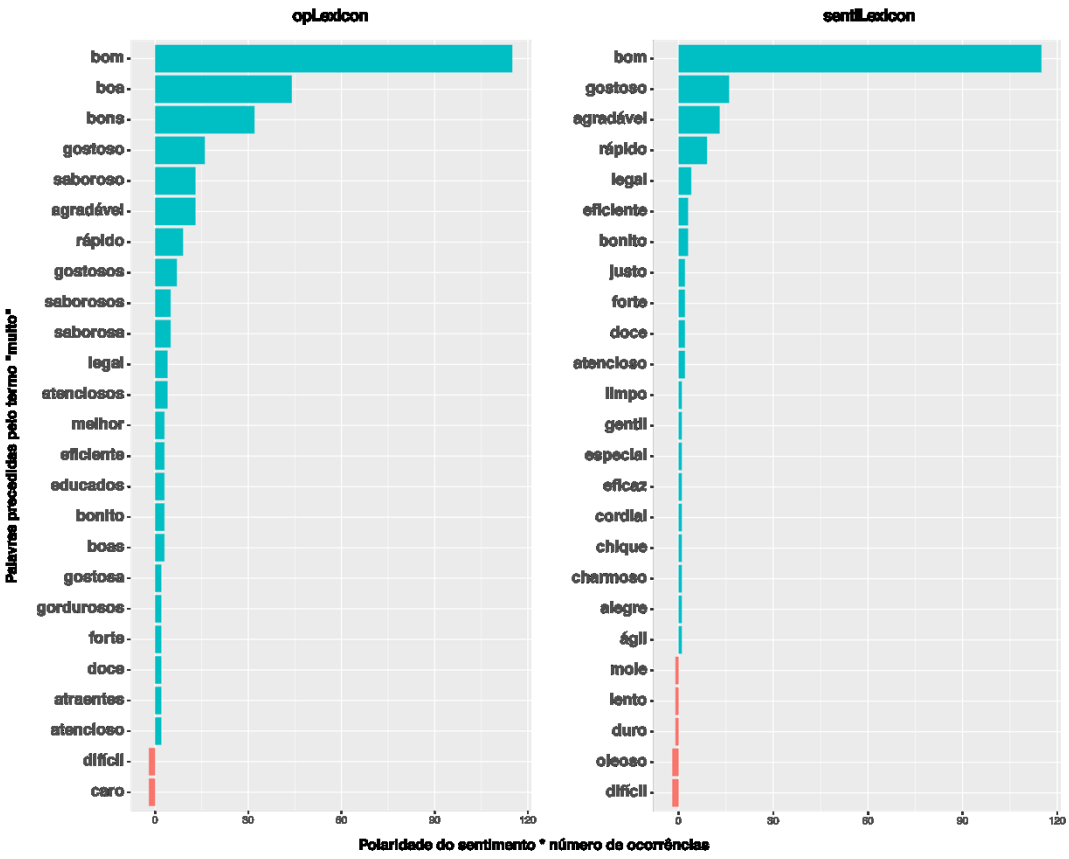
Figura 49 – Bigramas mais comuns em todo o *corpus* da EMPRESA 2



Fonte: Elaborado pelo autor.

Por algum motivo, a maionese e o atendimento parecem ser muito importantes para os clientes da EMPRESA 2. A próxima análise consiste no cruzamento dos bigramas com os léxicos, de onde se obtém a polaridade dos termos do *corpus* que constam dos léxicos. A Figura 50 mostra os termos mais frequentes precedidos pelo termo 'muito' considerando-se ambos os léxicos empregados.

Figura 50 – Termos mais frequentes precedidos pela palavra 'muito' da EMPRESA 2



Fonte: Elaborado pelo autor.

Novamente após o uso de ambos os léxicos chega-se a resultados semelhantes, mas com algumas diferenças a serem destacadas, como por exemplo os termos ‘muito + seco’, ‘muito + gorduroso’ e ‘muito + caro’, oriundos do opLexicon e dos termos ‘muito + decepcionado’ e ‘muito + ruim’, provenientes do sentiLex. Apesar destes termos possuírem baixa frequência se comparados a termos positivos, merecem ser avaliados dado seu teor crítico ao negócio.

A próxima análise objetiva a construção de uma rede de palavras. Neste processo nenhum dos léxicos foi usado, pois o único interesse é a relação entre os termos e a forma como estes constam do *corpus*. Conforme aplicação anterior optou-se por remover as *stop words* de ambos os conjuntos de termos, resultando apenas em bigramas que não são *stop words*, conforme apresentado no Quadro 35.

Quadro 35 –Dez bigramas mais frequentes no *corpus* sem *stop words* da EMPRESA 2

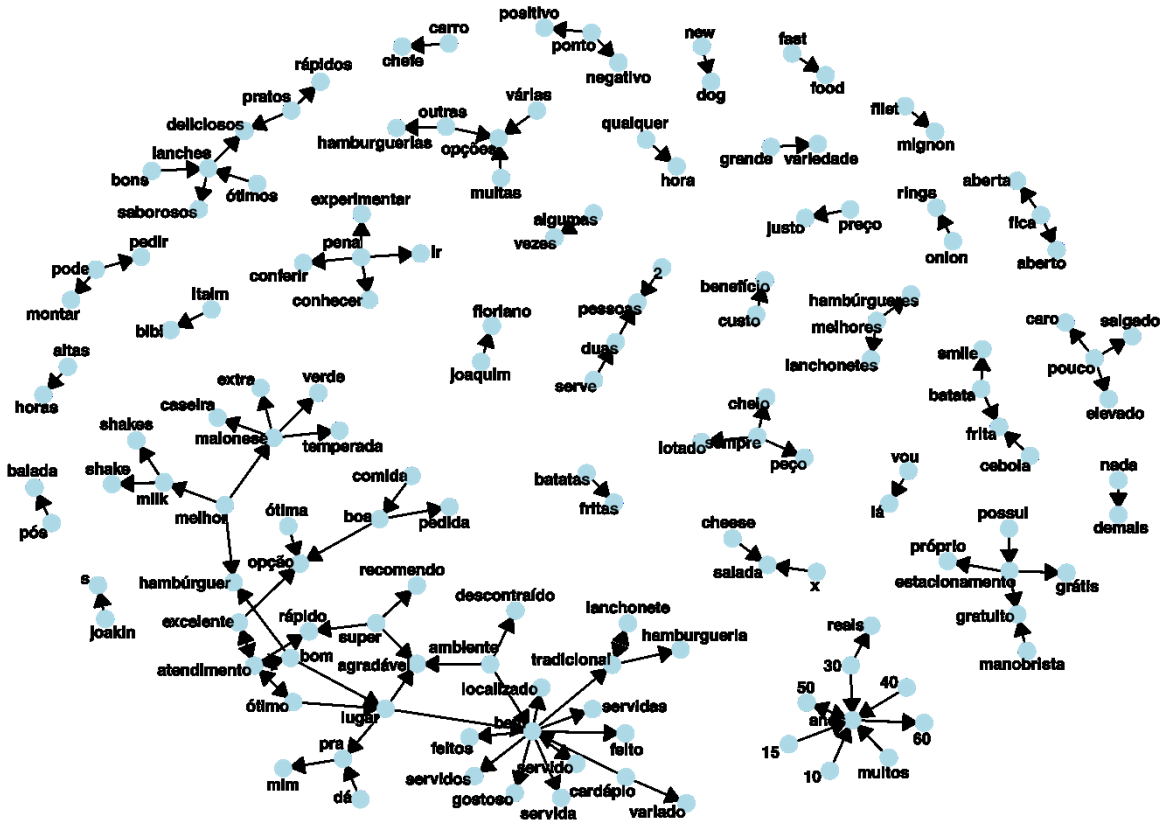
Termo1	Termo2	n
milk	shake	120
batata	frita	58
x	salada	47
bom	atendimento	39
joakin	s	36
cheese	salada	34
estacionamento	gratuito	33
ambiente	agradável	28
milk	shakes	27
preço	justo	27

Fonte: Elaborado pelo autor

Verifica-se novamente a presença de vários termos ricos para análise, uma vez que os clientes falam muito sobre a comida e sobre o lugar. Outro detalhe importante é o fato de os dez bigramas mais frequentes no *corpus* se referirem a aspectos positivos, o que se explica pela quantidade de opiniões positivas verificadas.

O último passo desta parte da análise é representar a rede de palavras oriunda dos dados apresentados no Quadro 35. Para a geração da rede de palavras foram filtrados os bigramas mais comuns baseados no critério de frequência (n > 5), conforme apresentado na Figura 51.

Figura 51 – Grafo de relações entre termos gerado a partir de bigramas da EMPRESA 2

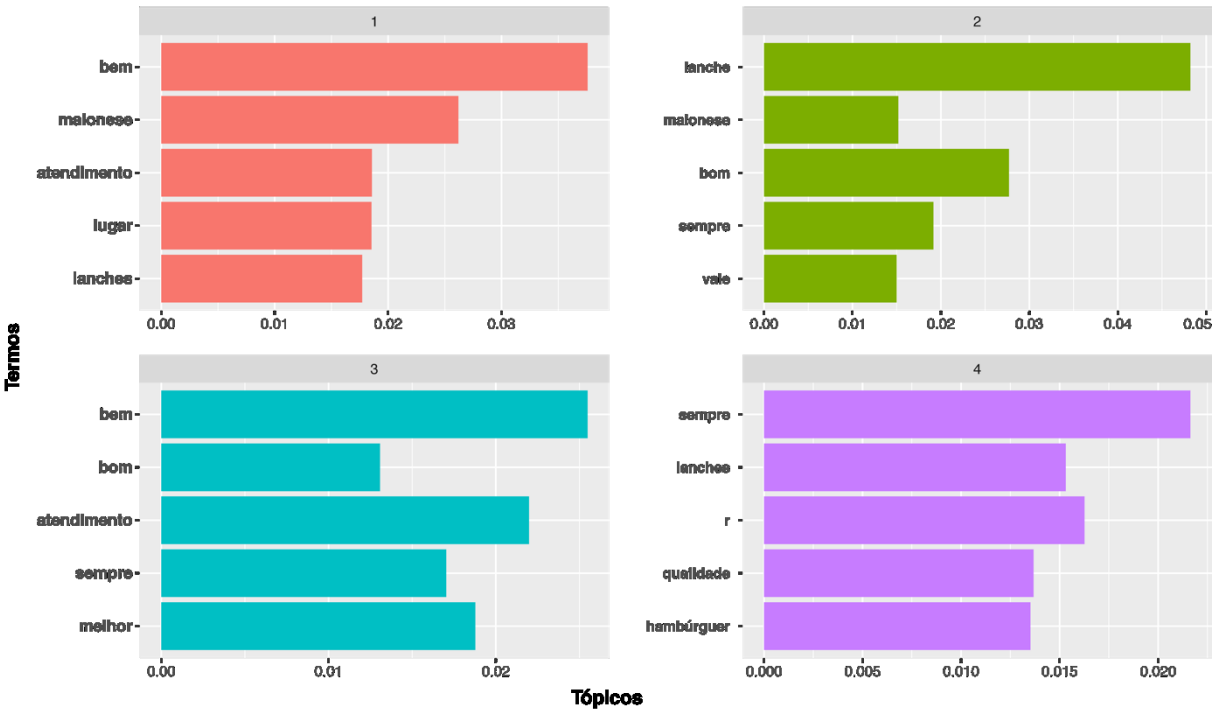


Fonte: Elaborado pelo autor.

A rede de palavras mostra a relação entre diversos termos interessantes para o domínio analisado nesta pesquisa, alguns positivos e outros negativos. Alguns exemplos são as relações entre termos como ‘sempre’, ‘cheio’, ‘lotado’ e ‘pouco + caro’, ‘pouco + salgado’ e ‘pouco + elevado’, que pode indicar pontos de atenção em relação à comida e ao preço praticado pelo estabelecimento.

Para finalizar a análise dos dados da EMPRESA 2, aplicou-se a técnica de Modelagem de Tópicos, tendo como parâmetro a geração de quatro tópicos com suas devidas distribuições de termos, o que pode ser visto na Figura 52.

Figura 52 – Resultado da Modelagem de Tópicos da EMPRESA 2



Fonte: Elaborado pelo autor.

Como a Modelagem de Tópicos não leva em consideração nenhum dos léxicos e avalia o *corpus* inteiro, a única operação de pré-processamento realizada durante a preparação dos dados foi a remoção de *stop words*.

Dada a distribuição encontrada pelo modelo, aparentemente o Tópico 1 fala sobre uma mistura de tópicos que envolve ‘comida’, ‘lugar’ e ‘atendimento’, o que se evidencia pela presença de termos como ‘atendimento’, ‘lanches’ e ‘lugar’. O Tópico 2 fala mais sobre comida, dada a presença de termos como ‘maionese’ e ‘lanches’. O Tópico 3 fala sobre ‘atendimento’ e o Tópico 4 sobre a ‘qualidade’ dos produtos.

Pode-se concluir, consideradas todas as análises elaboradas, que os clientes da EMPRESA 2 falam muito bem do restaurante. Em complemento, pôde-se verificar que quando os clientes comentam algo, lembram principalmente da comida, do lugar e do atendimento prestado pelo estabelecimento.

4.3 Análise dos dados da EMPRESA 3

A EMPRESA 3 é um restaurante especializado em hambúrgueres tradicionais americanos. Foi fundada por dois donos em 2012, que fizeram uma viagem aos EUA para conhecer receitas tradicionais de vários pontos do país. Na volta ao Brasil juntaram os melhores hambúrgueres que tinham provado e abriram o restaurante.

Além da presença no TripAdvisor, a EMPRESA 3 possui perfis no Facebook, contando com mais de 3.400 avaliações, e no Instagram, no qual possui mais de 16.900 seguidores. O restaurante está no TripAdvisor desde outubro de 2012.

Conforme processos já realizados nos experimentos anteriores, o processo de análise inicia-se com o pré-processamento dos dados, cujos resultados podem ser vistos no Quadro 36.

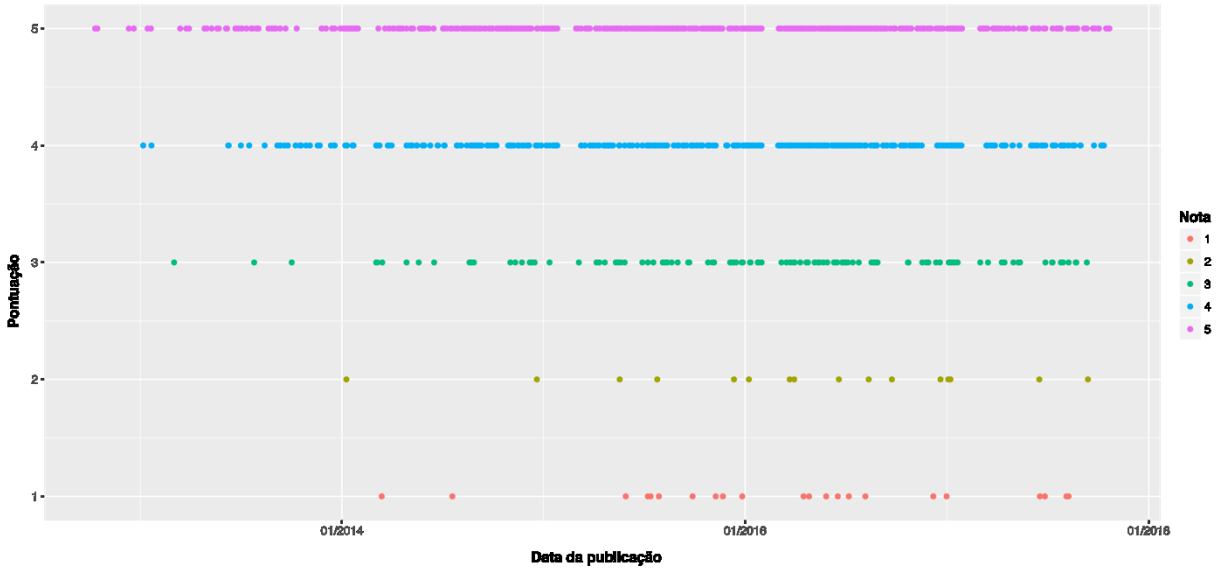
Quadro 36 – Sumarização dos dados da EMPRESA 3 após fase de pré-processamento

Procedimento	Quantidade de itens após procedimento
Entrada	1.275 documentos
Tokenização	58.668 palavras
Remoção de *stop words*	34.083 palavras
Remoção de números e caracteres especiais	33.716 palavras

Fonte: Elaborado pelo autor

Como resultado das etapas do pré-processamento, obteve-se um produto de 33.716 palavras sem *stop words*, distribuídas nos 1.275 documentos da EMPRESA 3. A análise dos dados começa com a visualização da distribuição das notas ao longo do tempo, apresentada na Figura 53.

Figura 53 – Distribuição das notas ao longo do tempo para a EMPRESA 3

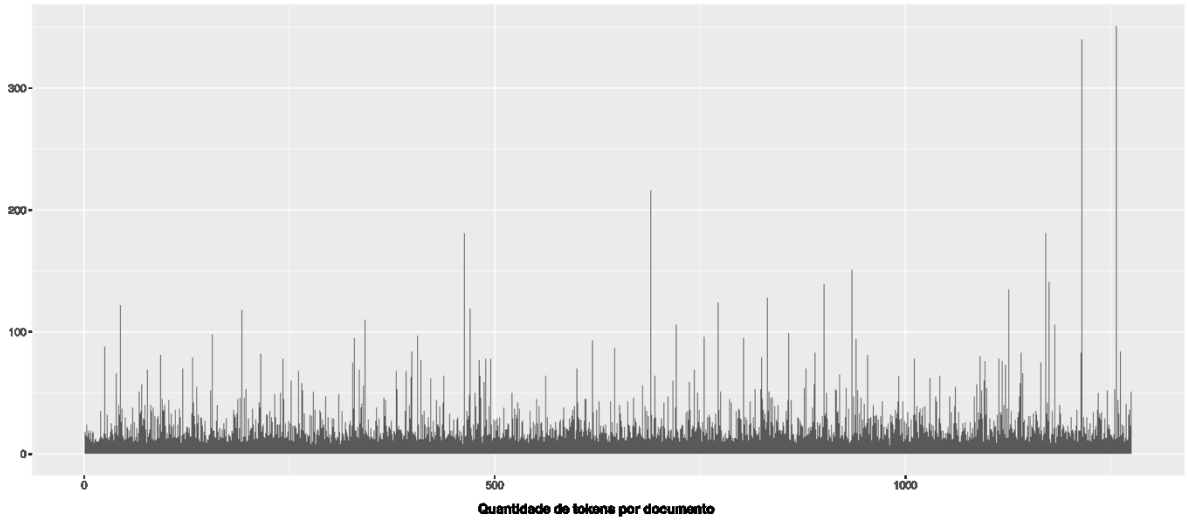


Fonte: Elaborado pelo autor.

Nota-se que a quantidade de notas positivas (4 e 5) supera muito as demais, o que confirma a predominância de opiniões positivas dos clientes sobre o restaurante. As análises serão aprofundadas visando obter uma melhor visão sobre o que os clientes falam quando fazem comentários positivos ou negativos sobre o estabelecimento.

A próxima análise representa o tamanho médio das avaliações por meio de um histograma que reflete a quantidade de termos por documento, conforme mostra a Figura 54.

Figura 54 – Quantidade de termos por documento da EMPRESA 3



Fonte: Elaborado pelo autor.

Analisando-se as repetições de palavras (as dez mais frequentes), nota-se que pela repetição absoluta dos termos no *corpus*, os clientes falam muito sobre a comida (1058[4] repetições), o local (739 repetições) e sobre o atendimento (442 repetições).

Quadro 37 – Dez termos mais presentes em todo o *corpus* da EMPRESA 3

Termo	Quantidade de repetições no *corpus*
bem	521
lanches	480
ambiente	468
atendimento	442
bom	409
hambúrguer	304
vale	275
lanche	274
lugar	271
melhor	268

Fonte: Elaborado pelo autor

O Quadro 38 considera os dez termos com mais repetições por documento, a fim de proporcionar uma visão diferente sobre a frequência de alguns termos.

Quadro 38 – Dez termos com mais repetições por documento da EMPRESA 3

Termo	Quantidade de repetições por documento
lugar	8
mesa	8
r	7
pra	6
burger	6
burger	6
burger	6
cheddar	5
hamburguer	5
bem	5

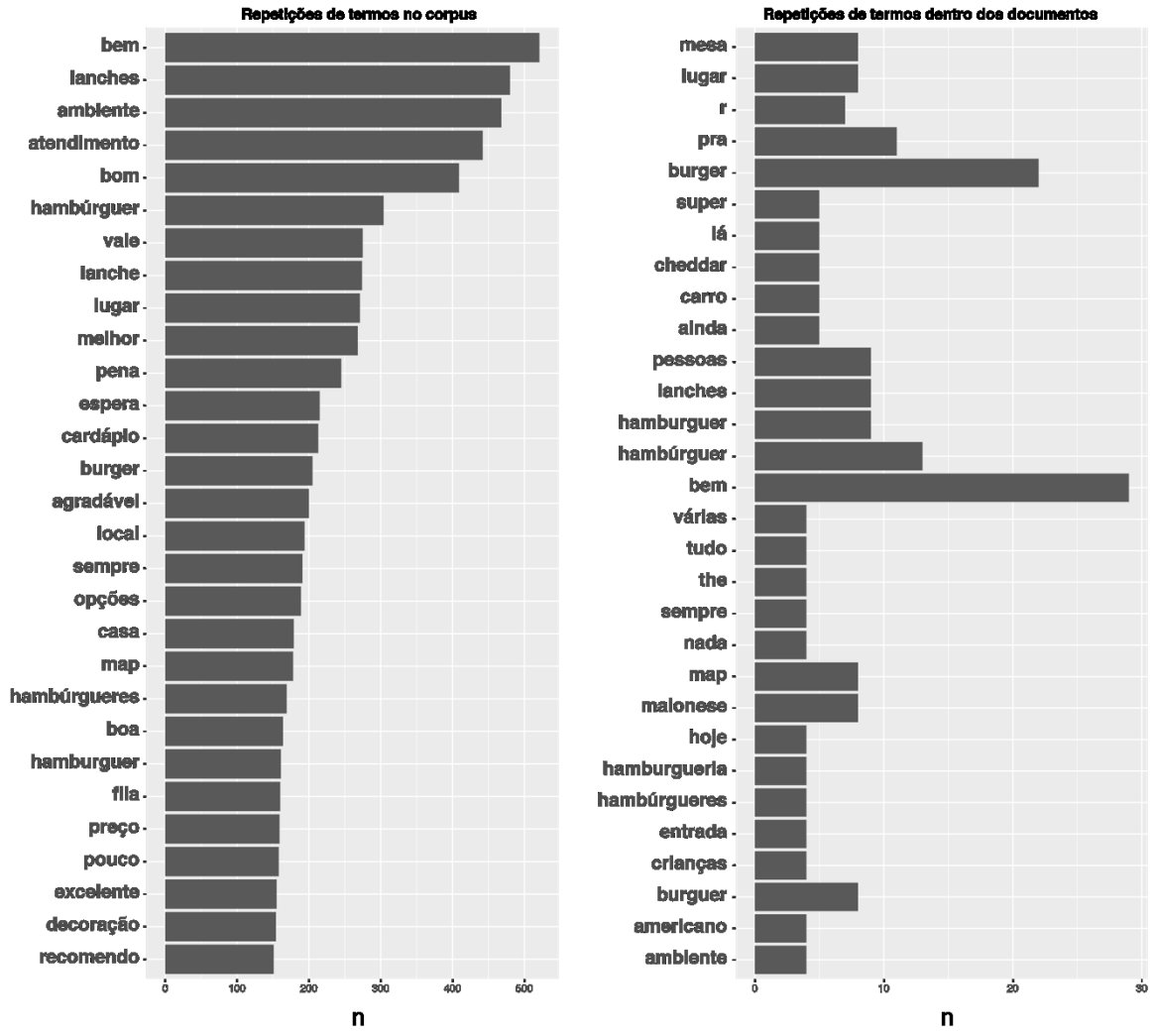
Fonte: Elaborado pelo autor

[4] Este número refere-se à soma dos termos que aparecem entre os dez mais citados.

Nota-se a repetição de termos como ‘burger’ e ‘hamburger’ e a presença de um termo alheio à análise (a letra ‘r’), que muito provavelmente refere-se ao termo ‘rua’, indicando a localização do restaurante, e que passou intacto pela remoção de *stop words*.

O gráfico apresentado na Figura 55 apresenta a plotagem de repetição de termos no *corpus* e dentro dos documentos.

Figura 55 – Repetição de termos no *corpus* e dentro dos documentos da EMPRESA 3

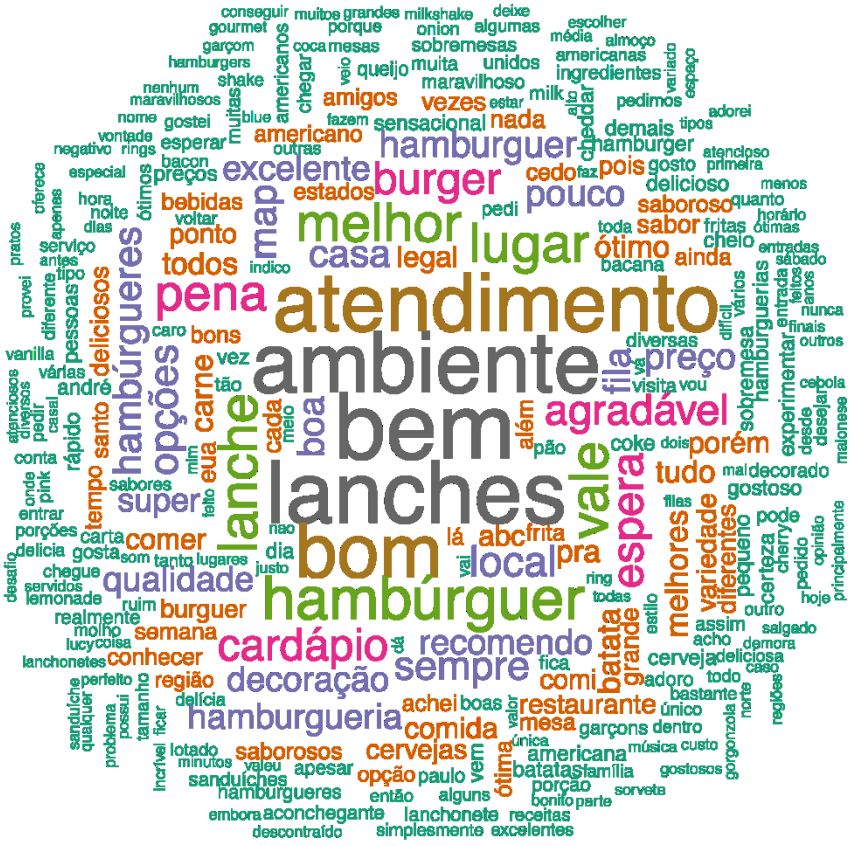


Fonte: Elaborado pelo autor.

Percebe-se, ao analisar-se o gráfico da Figura 55 que novamente as frequências são bem diferentes quando as observamos dentro dos documentos e em todo o *corpus*. O termo ‘bem’ aparece em ambos com enorme incidência. A partir daí, há apenas algumas repetições são consistentes.

A Figura 56 apresenta a nuvem de palavras que permite avaliar a presença de determinados termos em relação ao quanto estes termos se repetem no *corpus* da EMPRESA 3.

Figura 56 – Nuvem de termos mais frequentes da EMPRESA 3



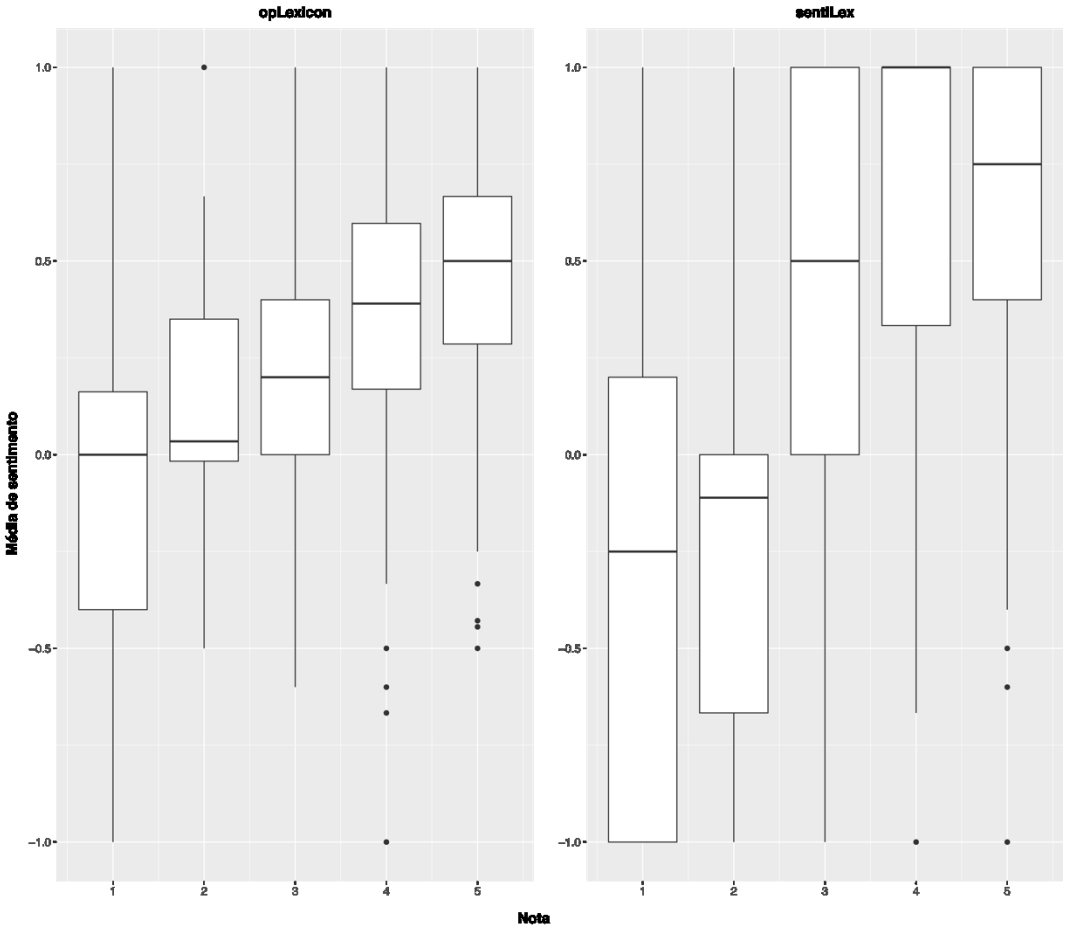
Fonte: Elaborado pelo autor.

A nuvem de palavras apresentada na Figura 56 indica que os termos mais frequentes neste *corpus* referem-se, majoritariamente, ao atendimento, ao lugar e à comida, dado que os mesmos possuem aparentemente a mesma proporção de tamanho. Os termos mais periféricos e que orbitam em torno dos centrais, falam sobre uma série de aspectos interessantes, incluindo itens como o 'cardápio' e a 'decoração', mas também revelam ocorrências que merecem uma atenção maior, como o termo 'espera', por exemplo. Dado seu tamanho, provavelmente é algo que se repete em vários documentos.

Procede-se agora à Análise de Sentimentos com o objetivo de avaliar as inclinações das postagens dos usuários. A primeira análise é que cruza os dados de pontuação atribuída pelo usuário numa escala de 1 a 5 e a classificação de sentimentos atribuída pelo léxico. A

Figura 57 apresenta um gráfico do tipo *box plot* que consiste na média de sentimento por nota do usuário da EMPRESA 3.

Figura 57 – Média de sentimento por avaliação segundo os léxicos para a EMPRESA 3

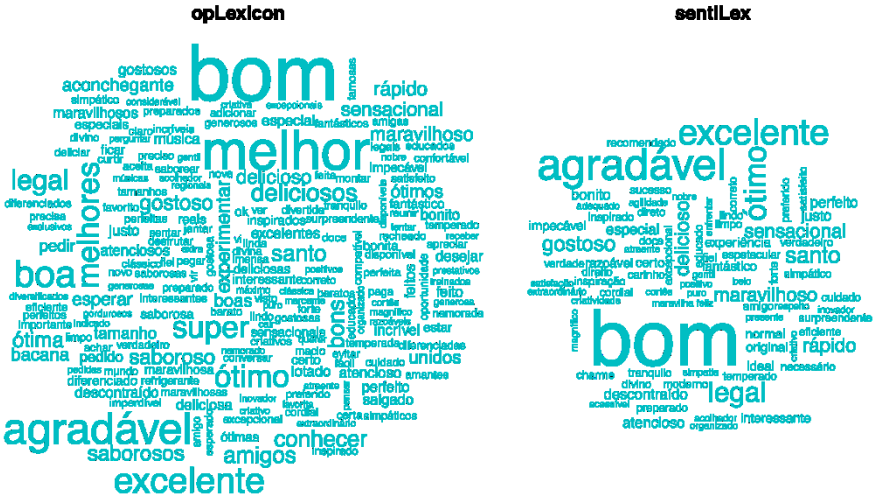


Fonte: Elaborado pelo autor.

Novamente constata-se a diferença proveniente da aplicação de ambos os léxicos no *corpus* da EMPRESA 3, com melhor distribuição para o léxico opLexicon. Ambos apresentam certa assimetria, mas o desiquilíbrio novamente é maior em relação ao léxico sentiLex. Neste caso a presença de *outliers* também é maior no léxico opLexicon.

A próxima análise foi realizada em relação à presença de termos positivos classificados por cada léxico e apresentados lado a lado na Figura 58.

Figura 58 – Nuvem de termos mais positivos por léxico para a EMPRESA 3

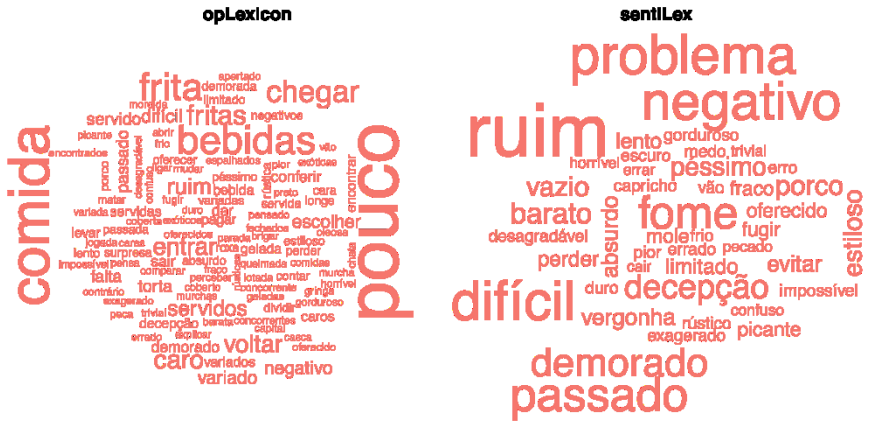


Fonte: Elaborado pelo autor.

Na nuvem de termos positivos gerada pelo cruzamento com o léxico opLexicon, os termos ‘bom’, ‘melhor’, ‘agradável’ e ‘excelente’ destacam-se. Já na nuvem de termos positivos gerada pelo cruzamento com o léxico sentiLex, os termos que mais se destacam foram ‘bom’, ‘agradável’ e ‘excelente’.

A Figura 59 apresenta a nuvem de termos negativos gerada pelo cruzamento com o léxico opLexicon, onde destacam-se os termos ‘comida’, ‘bebidas’ e ‘chegar’. Quanto à nuvem negativa de termos oriunda do cruzamento com o léxico sentiLex, destacam-se os termos ‘ruim’, ‘problema’, ‘negativo’, ‘difícil’, ‘demorado’ e ‘passado’, entre outros, como verificado na Figura 59.

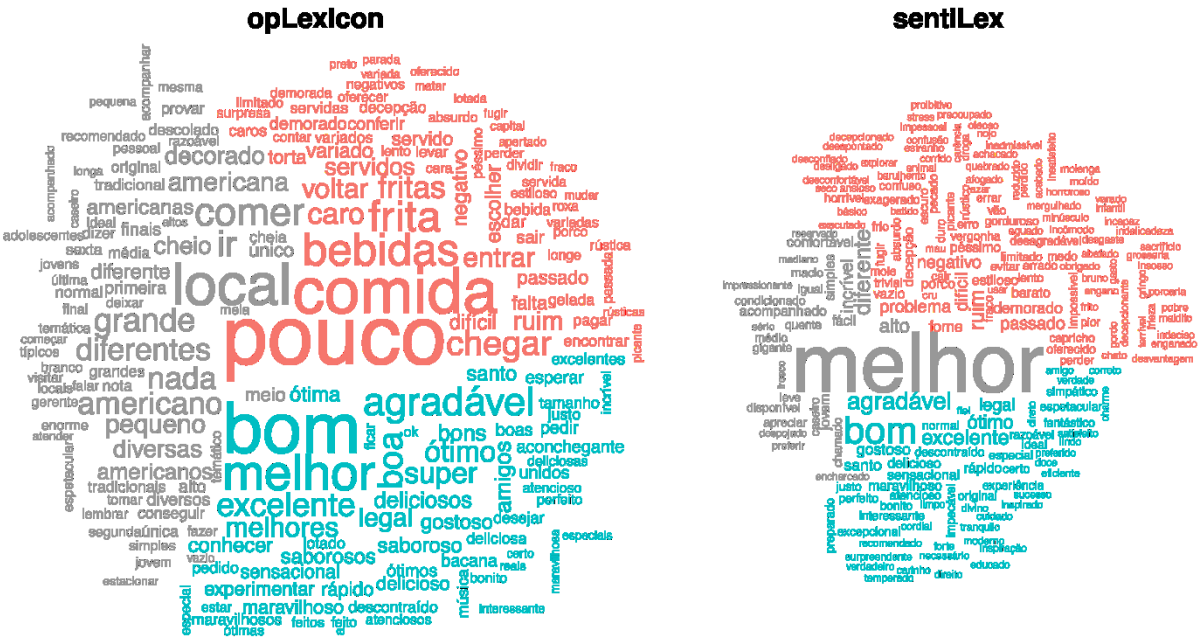
Figura 59 – Nuvem de termos negativos por léxico para a EMPRESA 3



Fonte: Elaborado pelo autor.

A Figura 60 apresenta a distribuição de termos com polaridade presentes no *corpus* considerando-se termos positivos, negativos e neutros.

Figura 60 – Nuvem de termos positivos, negativos e neutros da EMPRESA 3



Fonte: Elaborado pelo autor.

Considerando-se os termos presentes no léxico opLexicon, verifica-se que termos importantes para o domínio estudado, como por exemplo ‘comer, ‘local’ e ‘americano’ são considerados neutros, embora signifiquem muito no contexto em estudo, considerando-se a proposta do restaurante EMPRESA 3. Em relação ao léxico sentiLex, chama a atenção que o termo ‘melhor’ seja classificado como neutro e mais uma vez isto se repita também neste conjunto de dados.

A próxima análise consiste no cruzamento entre a relação dos termos positivos e negativos com a média das notas atribuídas pelos usuários, objetivando descobrir quais termos mais positivos e mais negativos estão associados com a média das avaliações conferidas pelos usuários.

O opLexicon foi o primeiro léxico a ser analisado desta forma, com os resultados expostos no Quadro 39.

Quadro 39 – Termos mais positivos e negativos com distribuições por documento e média de notas associadas segundo o léxico opLexicon para a EMPRESA 3

Termos mais positivos	Aparece em quantos documentos	Nota Média	Termos mais negativos	Aparece em quantos documentos	Nota Média
adequados	2	5.0	brigar	2	1.0
agradáveis	2	5.0	horrível	3	1.5
ampla	3	5.0	péssima	2	1.8
amplo	2	5.0	obrigado	2	2.0
animado	2	5.0	queda	2	2.2
ansioso	2	5.0	péssimo	8	2.4
apreciadores	2	5.0	barata	3	2.5
aproveitar	2	5.0	desconfortáveis	2	2.5
autêntico	2	5.0	minúsculo	2	2.5
bonitas	2	5.0	gorduroso	4	2.6

Fonte: Elaborado pelo autor

De forma similar, o resultado da classificação do léxico sentiLex é apresentado no Quadro 40, com destaque para os dez termos mais positivos e mais negativos.

Quadro 40 – Termos mais positivos e negativos com distribuições por documento e média de notas associadas segundo o léxico sentiLex para a EMPRESA 3

Termos mais positivos	Aparece em quantos documentos	Nota Média	Termos mais negativos	Aparece em quantos documentos	Nota Média
animado	2	5.0	horrível	3	1.0
ansioso	2	5.0	obrigado	2	1.5
autêntico	2	5.0	péssimo	8	1.8
digno	2	5.0	minúsculo	2	2.0
errar	3	5.0	gorduroso	4	2.2
erro	4	5.0	decepção	10	2.4
executado	2	5.0	decepcionante	2	2.5
fenomenal	2	5.0	pior	4	2.5
fresco	3	5.0	usar	2	2.5
impressionante	2	5.0	mediano	3	2.6

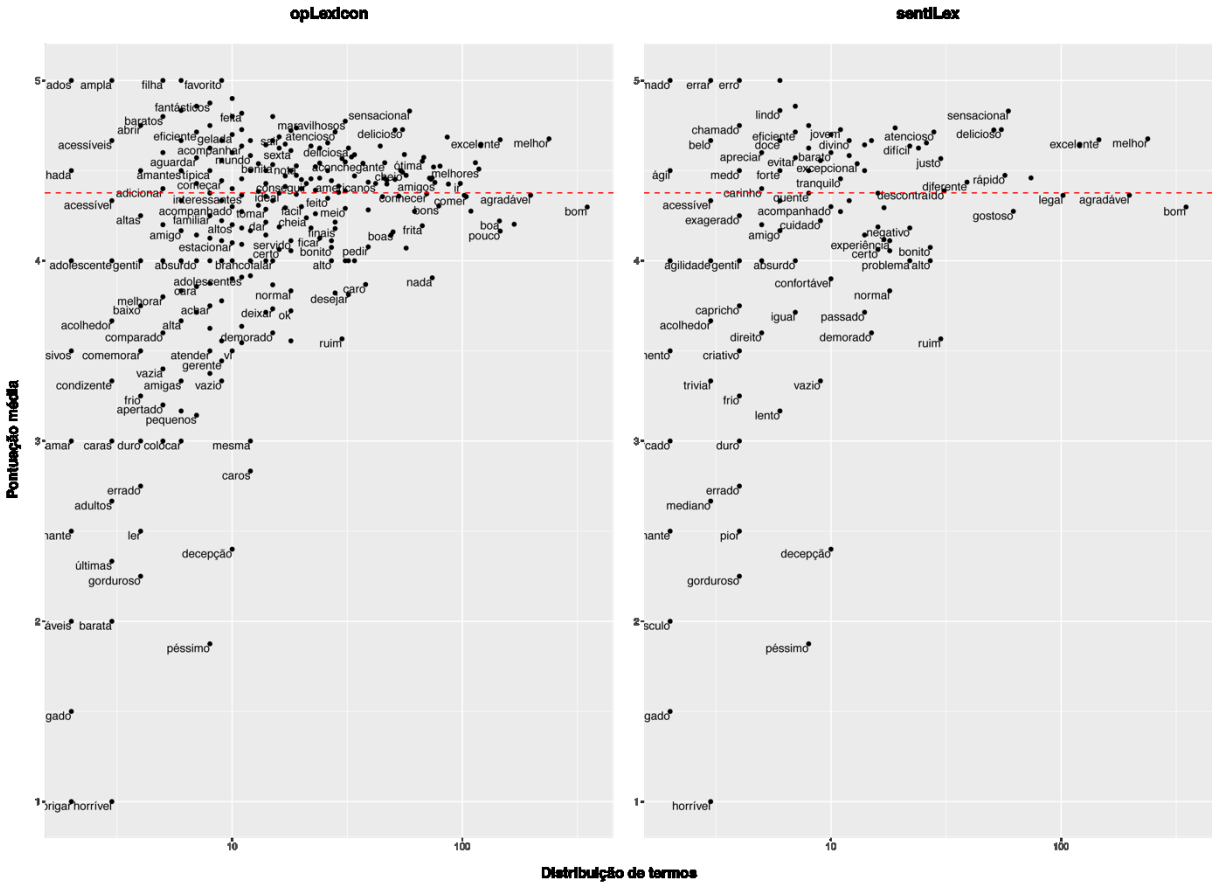
Fonte: Elaborado pelo autor

Ao comparar-se os resultados do Quadro 39 e Quadro 40 observa-se que os termos classificados pelo léxico opLexicon parecem fazer mais sentido do que alguns resultados classificados pelo sentiLex, com destaque para os termos ‘erro’ e ‘errar’ que aparecem na

análise do sentiLex como termos positivos. Uma outra observação importante é a aparição do termo ‘péssimo’ que aparece na classificação do opLexicon e o termo ‘decepção’, que surge da classificação do sentiLex. Talvez indiquem experiências muito negativas dos clientes, devendo ser analisados com maior atenção.

Avaliando-se a distribuição dos termos em relação à pontuação média, é possível visualizar melhor diferenças entre as classificações feitas por ambos os léxicos, conforme mostra a Figura 61.

Figura 61 – Distribuição de termos por avaliação segundo os léxicos para a EMPRESA 3

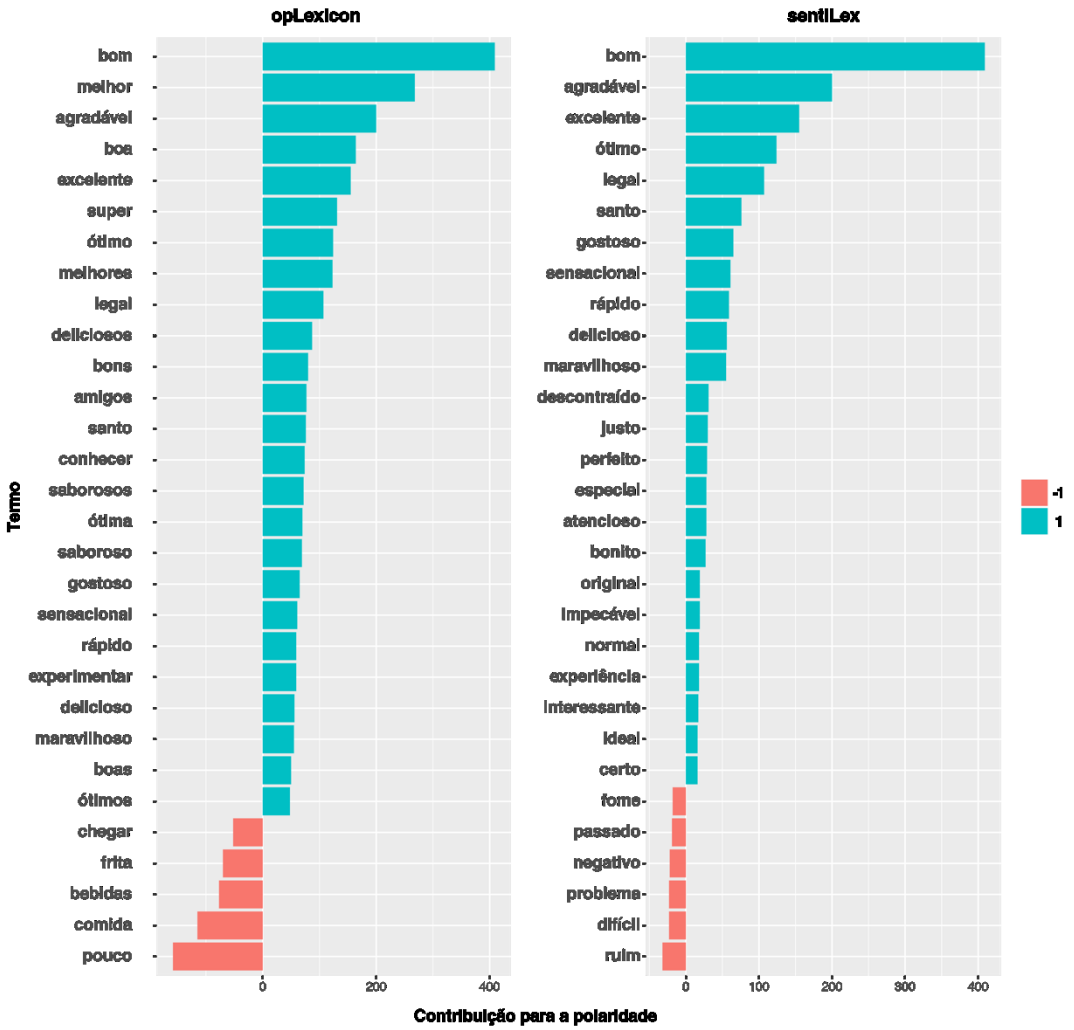


Fonte: Elaborado pelo autor.

Comparando-se os gráficos, nota-se novamente a presença de termos comuns a ambos os léxicos, tais como ‘agradável’, ‘excelente’ e ‘bom’. Entretanto, o opLexicon parece descrever melhor o que se passa com o conjunto de dados em função da riqueza e quantidade de termos classificados.

Para fins de comparação, a Figura 62 mostra a frequência de termos positivos e negativos em oposição, classificados por ambos os léxicos para a EMPRESA 3.

Figura 62 – Termos positivos e negativos mais presentes no *corpus* da EMPRESA 3

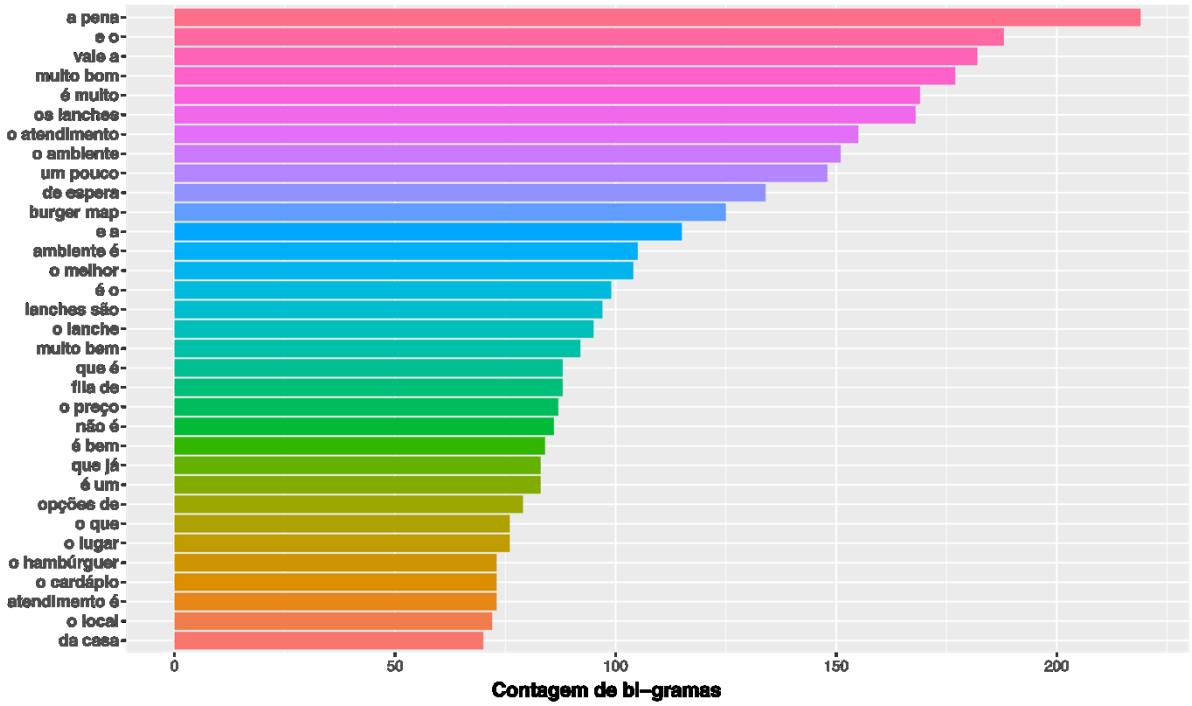


Fonte: Elaborado pelo autor.

O gráfico da Figura 62 mostra que, desta vez, poucos termos coincidem, como os positivos ‘bom’, ‘agradável’ e ‘excelente’. Porém, alguns termos negativos merecem atenção, como os termos ‘chegar’, ‘problema’ e ‘difícil’, que se repetem algumas vezes.

Visando entender melhor a relação entre certos termos que aparecem frequentemente juntos, as próximas análises consideram bigramas, a começar pela plotagem de termos que mais aparecem juntos em todo o *corpus*, como mostra a Figura 63.

Figura 63 – Bigramas mais comuns em todo o *corpus* da EMPRESA 3

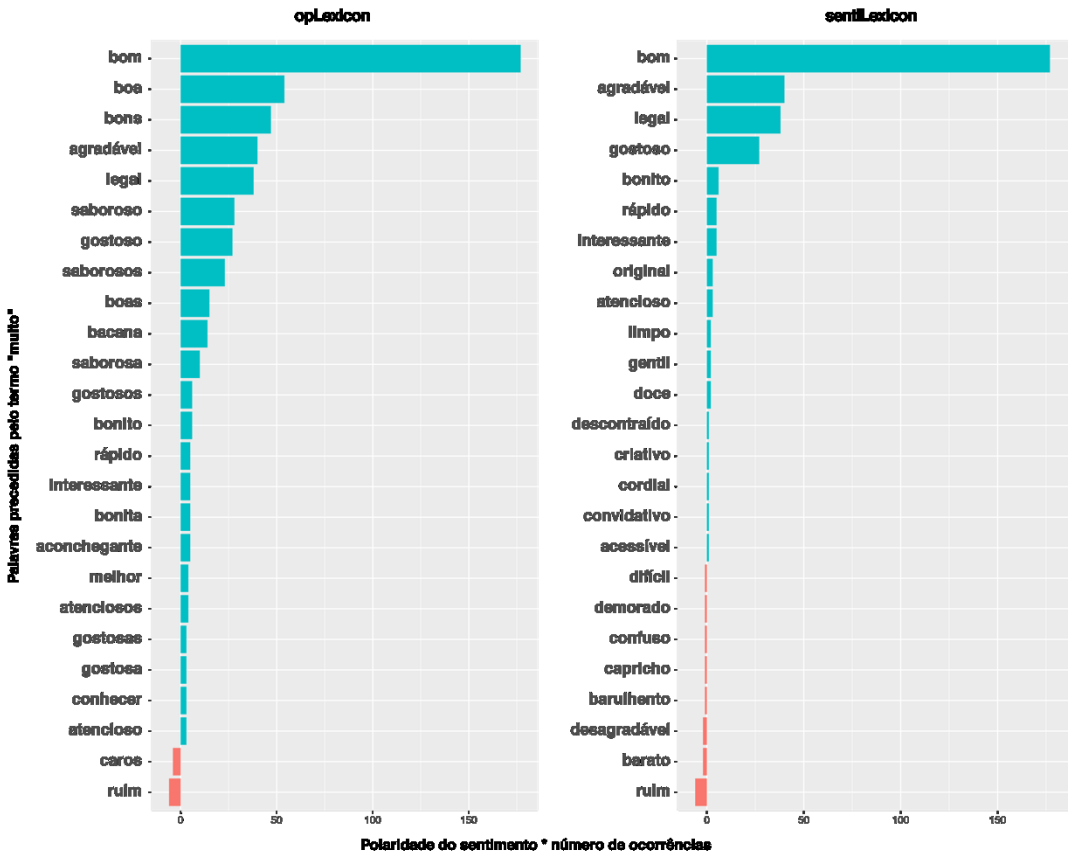


Fonte: Elaborado pelo autor.

Pelos bigramas apresentados na Figura 63, os clientes da EMPRESA 3 falam muito sobre os três termos mais comuns em todas as análises: a comida, o lugar e o atendimento. Entretanto, a análise de bigramas, ainda que em seu início, começa a trazer novas perspectivas sobre termos que apareceram em várias das análises feitas até agora, a exemplo do que se fala sobre ‘espera’ e ‘fila’, e que poderá ser evidenciado em outras análises.

A próxima análise consiste no cruzamento dos bigramas com os léxicos, pela qual se obtém a polaridade dos termos do *corpus* que constam dos léxicos. A Figura 64 mostra os termos mais frequentes precedidos pelo termo 'muito', considerando ambos os léxicos.

Figura 64 – Termos mais frequentes precedidos pela palavra 'muito' da EMPRESA 3



Fonte: Elaborado pelo autor.

Novamente após o uso de ambos os léxicos chegou-se a resultados semelhantes, mas com algumas diferenças notáveis, como por exemplo os termos ‘muito + caros’ e ‘muito + ruim’, oriundos do léxico opLexicon e os termos ‘muito + desagradável’, ‘muito + barulhento’, ‘muito + demorado’ e ‘muito + confuso’, provenientes do sentiLex. Apesar destes termos possuírem baixa frequência se comparados aos termos positivos identificados, merecem ser observados.

A próxima análise objetiva verificar a relação entre os termos e a forma como estes constam do *corpus*. Conforme aplicação anterior, foram removidas apenas as *stop words* de ambos os conjuntos de termos, resultando em bigramas que não são *stop words*, conforme apresentado no Quadro 41.

Quadro 41 – Dez bigramas mais frequentes no *corpus* sem *stop words* da EMPRESA 3

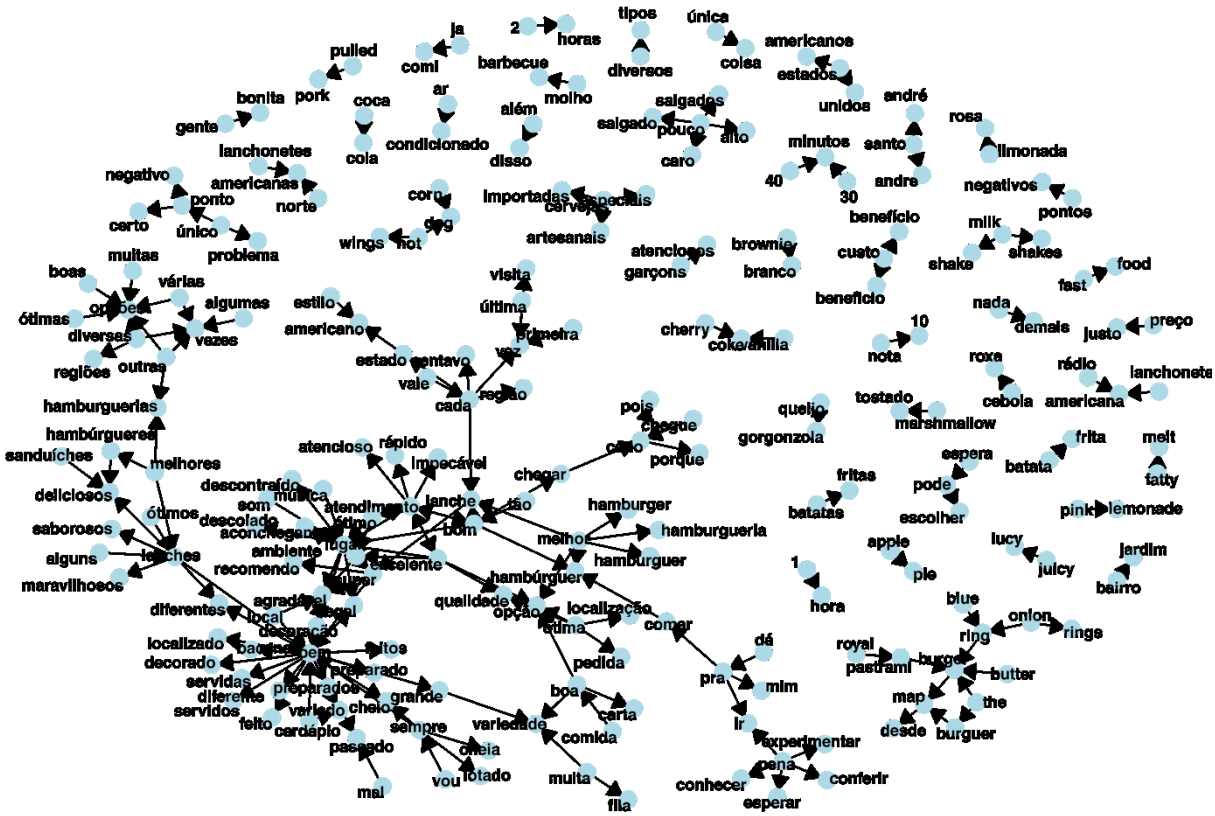
Termo1	Termo2	n
burger	map	125
santo	andré	64
batata	frita	62
bom	atendimento	53
bem	decorado	50
ambiente	agradável	49
burguer	map	48
estados	unidos	48
the	burger	43
melhor	hamburgueria	37

Fonte: Elaborado pelo autor

Ao analisar-se os bigramas apresentados no Quadro 41 observa-se que os clientes falam muito sobre a comida, sobre o lugar e sobre o atendimento, os três aspectos mais frequentes em toda a pesquisa. Destaca-se que 'batata frita' seja o termo mais comentado que o principal prato oferecido pelo estabelecimento: hambúrguer. Importante notar que, mais uma vez, os dez bigramas mais frequentes no *corpus* referem-se a aspectos positivos, o que se explica pela quantidade de opiniões positivas em relação às opiniões negativas.

O último passo desta parte da análise é representar a rede de palavras oriunda dos dados apresentados no Quadro 41. Para a geração da rede de palavras foram filtrados os bigramas mais comuns baseados no critério de frequência ($n > 5$), conforme apresentado na Figura 65.

Figura 65 – Grafo de relações entre termos gerado a partir de bigramas da EMPRESA 3

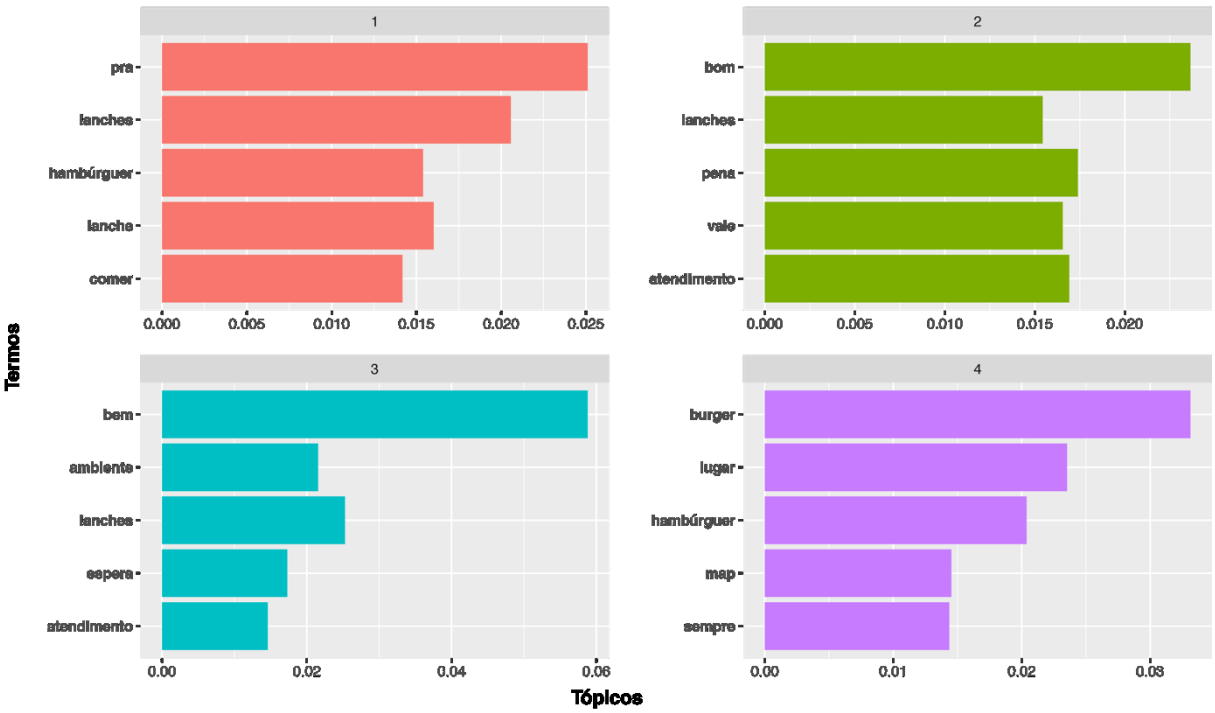


Fonte: Elaborado pelo autor.

A rede de palavras mostra a relação entre diversos termos interessantes para o domínio analisado nesta pesquisa, alguns positivos e outros negativos. Alguns exemplos são as relações entre termos como ‘muita’ e ‘fila’ e ‘2’ e ‘horas’, como aspectos possivelmente negativos relacionados ao tempo de espera. Em contraponto, nota-se também termos como ‘vale + cada + centavo’ e ‘garçons + atenciosos’, que denotam o quanto as pessoas gostam do lugar e do atendimento.

Para finalizar a análise dos dados da EMPRESA 3, aplicou-se a técnica de Modelagem de Tópicos com os mesmos parâmetros usados até agora, o que pode ser visto na Figura 66.

Figura 66 – Resultado da Modelagem de Tópicos da EMPRESA 3



Fonte: Elaborado pelo autor.

Dada a distribuição encontrada pelo modelo, aparentemente o Tópico 1 fala sobre ‘comida’, o que se evidencia pela presença dos termos ‘lanches’, ‘lanche’ e ‘hamburguer’. O Tópico 2 possivelmente fala sobre a ‘experiência’, dada a presença de termos como ‘lanches’, ‘atendimento’, ‘vale’ e ‘pena’. O Tópico 3 fala sobre uma mistura de tópicos que tem a ver com ‘atendimento’ e ‘lugar’, considerando-se a presença dos termos ‘atendimento’, ‘lanche’ e ‘espera’. Já o Tópico 4 fala sobre o ‘lugar’ e sobre a ‘comida’, considerando-se a ocorrência de termos como ‘lugar’, ‘burger’ e ‘hamburguer’.

Pode-se concluir, consideradas todas as análises efetuadas, que os clientes da EMPRESA 3 falam muito bem do restaurante. Quando opinam sobre ele, lembram principalmente sobre a comida, o lugar e o atendimento, os três aspectos mais constantes em todas as análises realizadas.

4.4 Considerações sobre as análises dos dados das empresas

As análises desenvolvidas por meio do *framework* apresentado nos dados de três empresas do segmento de alimentação revelaram diversos aspectos interessantes sobre as

opiniões destes clientes. Há de se ressaltar que tal análise (e a evidenciação dos resultados alcançados), se realizadas manualmente, dificilmente trariam a mesma riqueza de informações oriunda das técnicas empregadas no *framework* desenvolvido nesta pesquisa.

As análises procedidas foram realizadas de forma metódica por meio da aplicação de um *Framework* para Mineração de Opiniões cuja construção foi guiada pelos princípios do *tidy data*, conceito que preconiza a estruturação de dados de tal forma a facilitar todos os processos de mineração aos quais estes dados venham a se submeter (WICKHAM, 2014; WICKHAM; GROLEMUND, 2016; SILGE; ROBINSON, 2017).

Os efeitos da adoção desta abordagem para trabalhar dados, sobretudo dados não estruturados como é o caso do domínio considerado nesta pesquisa, foram evidenciados nos procedimentos realizados para esta pesquisa, dado que a abordagem foi aplicada desde a extração dos dados até a geração de visualizações gráficas dos resultados, comprovando sua eficiência em aplicações de Mineração de Texto.

Trabalhar com a lógica de um *token* por linha (SILGE; ROBINSON, 2017) possibilitou com que todas as análises fossem realizadas de forma muito rápida e organizada, o que, por sua vez, proporciona maior flexibilidade, evidenciada nesta pesquisa pela adição de técnicas de Análise de Sentimentos e Modelagem de Tópicos ao fluxo de mineração de textos clássico (FELDMAN; SANGER, 2007).

O *framework* proposto e empregado nesta pesquisa provou-se capaz de cumprir os objetivos específicos propostos, sobretudo considerando-se as limitações impostas pela natureza da pesquisa, que começou pela delimitação do universo de dados a serem trabalhados. Considerando-se o fato de as empresas que constituíram o objeto desta pesquisa sejam pequenas e médias, esperava-se que houvesse tal limitação em relação aos dados.

Não obstante, esta limitação implicou em escolhas que nortearam todo o estudo efetuado, tais como o emprego da técnica de mineração de textos baseada em léxico e outras decisões de menor impacto para os resultados, tal como a não adoção de *stemming* durante o pré-processamento dos dados.

Demonstrou-se, portanto, que a Mineração de Textos tem potencial para revelar padrões (FELDMAN; SANGER, 2007) e possibilita a busca de conhecimento. Assim, o *framework* empregado pode ser usado para compreender melhor o cliente, suas expectativas e até mesmo suas frustrações, gerando assim conhecimento acerca dos clientes para benefício da empresa. A literatura afirma que “a partir de uma melhor compreensão do cliente, a empresa terá maior compreensão das verdadeiras necessidades e expectativas deste cliente” (GARCÍA-MURILLO; ANNABI, 2002, p. 882).

Estas análises revelaram, entre outras coisas, que os aspectos mais abordados pelos clientes referem-se à comida, ao lugar e o atendimento, variando em intensidade e polaridade, não obstante, quase sempre de forma positiva. Isto talvez se deva ao fato de as opiniões dos usuários serem, em sua grande maioria focadas na experiência como um todo, e não apenas em um ou poucos aspectos específicos. Tal descoberta consiste, em última instancia, em conhecimento útil sobre o comportamento e inclinações dos clientes que frequentam os restaurantes abordados nesta pesquisa.

Nota-se em todos os conjuntos de dados estudados que algumas das opiniões são bem densas, enquanto outras são mais sucintas, mas, ainda assim, aplicando-se técnicas de Mineração de Textos foi possível ter uma ideia da frequência com que quanto determinados termos se repetem e o quanto estes influenciam no resultado final das análises. Neste ponto, há de se enfatizar que apenas contar o quanto um termo se repete em um *corpus* ou dentro de documentos proporciona um panorama interessante, mas que precisa ser confirmado ou refutado por análises mais profundas. Assim sendo, a Análise de Sentimentos cumpriu um papel importante para a geração de conhecimento acerca do cliente das empresas analisadas.

Saber sobre o que as pessoas mais falam é apenas o primeiro passo. Em se tratando de informação relevante, sobretudo para as empresas, faz-se necessário saber o quão bem ou mal as pessoas falam quando escrevem sobre qualquer aspecto de sua experiência num restaurante que frequentam.

Pang e Lee (2008) afirmam que as opiniões de outras pessoas sempre foram relevantes para que as pessoas decidam tomar decisões em relação a ter ou não contato com determinada experiência. Considerando-se a popularização da internet e, sobretudo das tecnologias móveis, sites que recebem opiniões de pessoas sobre produtos e serviços tornaram-se muito populares nos últimos anos.

Nesta pesquisa optou-se pela rede social TripAdvisor, especializada em receber opiniões de usuários sobre suas experiências com empreendimentos como hotéis, restaurantes e eventos. O foco desta pesquisa voltou-se especificamente aos restaurantes, tendo abordado três estabelecimentos especializados no mesmo tipo de produto: hambúrgueres.

As análises de sentimento nesta pesquisa foram baseadas em léxicos, com a aplicação de dois únicos léxicos validados na literatura em português do Brasil: o léxico opLexicon versão 3.0 (SOUZA; VIEIRA, 2012) e o léxico sentiLex (SILVA *et al.*, 2010).

Considerando-se as limitações dos léxicos empregados, ambos apresentaram resultados que podem ser considerados bons, dependendo da ótica de análise enfocada. Entretanto, convém considerar que nenhum dos dois léxicos foi criado para ser usado no

domínio abordado nesta pesquisa (restaurantes do tipo hamburgueria), o que resultou em resultados por vezes peculiares, conforme apontado em vários momentos das análises expostas, como por exemplo, a atribuição de polaridade negativa a termos aparentemente positivos, considerando-se a natureza dos dados e o domínio pesquisado.

Segundo Pang e Lee (2008), o sentimento e a subjetividade são bastante sensíveis ao contexto e, de certa forma, dependentes do domínio, mesmo considerando-se que a noção geral de opiniões positivas e negativas é bastante consistente em diferentes domínios. Talvez por isso, esta pesquisa tenha chegado a resultados interessantes, mesmo considerando-se que os léxicos classificassem termos importantes de forma equivocada.

Não obstante, consideradas todas as análises e comparando-se a maneira como ambos os léxicos classificaram os termos analisados, observou-se que o léxico opLexicon versão 3.0 (SOUZA; VIEIRA, 2012) teve maior consistência nos resultados. Isto deve-se ao fato de o léxico opLexicon ser melhor estruturado e muito maior que o léxico sentiLex, o que possibilitou uma maior abrangência de termos classificados. Tal resultado, por sua vez, reflete diretamente na qualidade das análises, denotada pelos gráficos das análises dos dois léxicos, nos quais ambos aparecem lado a lado.

Outro fator interessante a abordar tem relação com os processos de pré-processamento, que antecedem os processos de mineração de textos em si (FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2017). Nesta pesquisa algumas decisões relacionadas ao pré-processamento dos textos foram tomadas, conforme explicado no tópico ‘Coleta e pré-processamento dos dados’. Entre estas decisões destacam-se a remoção de *stop words* e o uso de *stemmer*.

A literatura afirma que a remoção de *stop words* é necessária, dado o fato de que os termos que se repetem e não carregam em si nenhum sentido semântico devem ser removidos preliminarmente, visando garantir que o *corpus* não contenha ruídos quando submetido aos processos subsequentes de mineração de textos (FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2017), o que foi aplicado nesta pesquisa em quase todos os processos. A única exceção refere-se à parte das análises que envolveu bigramas (FELDMAN; SANGER, 2007; SILGE; ROBINSON, 2017).

Esta decisão tem a ver com o fato de que algumas análises visavam descobrir termos que vinham juntos de outros, considerados *stop words* como, por exemplo, o termo ‘muito’. Esta análise permitiu descobrir combinações de termos frequentemente usadas pelos clientes contendo expressões mais abrangentes, tais como ‘muito bom’ ou ‘muito ruim’, além de

aspectos mais específicos sobre algum serviço ou produto oferecido pelos restaurantes, a exemplo de ‘muito demorado’ ou ‘muito caro’.

Por uma questão de uniformidade das análises e rigor do método selecionado, a mesma lista de *stop words* foi adotada em todas as análises, embora a literatura sugira que a lista de *stop words* possa (e até mesmo deva) ser customizada considerando-se o contexto e objetivo da mineração de textos (SILVA; PERES; BOSCARIOLI, 2017).

Outra decisão tomada na seleção do método empregado por esta pesquisa refere-se ao emprego de *stemming*, técnica que permite a redução de termos ao seu radical (SILVA; PERES; BOSCARIOLI, 2017). Optou-se nesta pesquisa pela não utilização de *stemming* por dois motivos, conforme já indicados no capítulo de procedimentos metodológicos: o primeiro é o fato de o tamanho da massa de dados analisada em cada caso ser demasiado pequeno, o que implica que reduzir termos talvez não seja uma decisão sábia. Em segundo lugar, a eficiência e precisão dos *stemmers* em língua portuguesa é muito baixa (SILVA; PERES; BOSCARIOLI, 2017), sem contar a falta de desenvolvimento de *stemmers* para o português desde 2009, conforme apontado por Singh e Gupta (2017).

Em complemento, há de se considerar ainda o problema de redução dos termos na língua portuguesa, uma vez que esta possui complexidade maior que outras línguas, como é o caso do inglês devido, sobretudo, ao grande número de formas para o plural e inúmeras flexões verbais, além das diferentes regras para tratamento de aumentativos, diminutivos e gêneros (SILVA; PERES; BOSCARIOLI, 2017). Assim, considerando-se os motivos expostos, justifica-se a decisão de não aplicação da técnica de *stemming* nesta pesquisa.

Ainda sobre a escolha do uso de bigramas, convém ressaltar que constatou-se uma grande riqueza de análises que só foi possível graças ao estudo dos termos que aparecem lado a lado no *corpus* prospectado nesta pesquisa. Parte considerável das análises executadas nesta pesquisa levam em consideração unigramas, tendo proporcionado diversas visualizações de dados e bons resultados. Entretanto, ao analisar-se a combinação de termos que aparecem frequentemente no *corpus* enfocado, foi possível descobrir fatos e aspectos interessantes que, se vistos como unidades separadas, não trariam o mesmo sentido daquele descoberto quando considerados os bigramas analisados.

A título de exemplo, destaca-se nas análises os termos ‘muito’ e ‘demorado’. Se vistos de forma independente, ambos podem significar coisas diferentes, mas se vistos de forma agregada a outros termos, podem significar algo que incomoda os clientes, como a demora no atendimento, demora numa fila ou até mesmo demora para que o pedido chegasse à mesa. Os três aspectos indicados nos exemplos expostos são negativos para a experiência dos clientes,

podendo impactar em outros aspectos do restaurante, muito embora tenha sido constatado que grande parte das opiniões cedidas pelos clientes trate de fatores positivos. Assim sendo, ao se empregar análises que considerem bigramas, foi possível descobrir diversas expressões positivas e algumas negativas que precisam ser analisadas com mais cuidado, graças sobretudo, ao impacto que estas podem exercer sobre outros clientes do restaurante também participantes de sua rede social.

Pang e Lee (2008) citam dados de uma pesquisa que contou com mais de 2000 participantes afirmando que entre os leitores de avaliações on-line de restaurantes, hotéis e vários tipos de serviços (como por exemplo, agências de viagens ou médicos), entre 73% e 87% relataram que as avaliações online tiveram uma influência significativa em sua decisão de compra.

Quando termos como ‘muito bom’ ou ‘nota 10’ aparecem associados a opiniões positivas, há chances aumentadas de que os clientes em potencial tornem-se clientes de fato. Da mesma forma que expressões como ‘muito ruim’ ou ‘super demorado’ aparecem com certa frequência no *corpus* analisado, pode haver um impacto negativo na intenção destas pessoas tornarem-se clientes de fato do restaurante.

De certa forma, a Análise de Sentimentos é uma maneira de alcançar conhecimento sobre o cliente, que por sua vez é citado na literatura como uma forma de obter uma sensação do sentimento geral e inclinações dos clientes. Isto, por sua vez, pode ajudar as empresas no gerenciamento de crises e em possíveis melhorias do negócio como um todo (SALOMANN *et al.*, 2005; CASTELLANOS *et al.*, 2011).

Outro aspecto interessante do *framework* delineado nesta pesquisa refere-se à aplicação da técnica de Modelagem de Tópicos (BLEI; NG; JORDAN, 2003; BLEI; LAFFERTY, 2009; BLEI, 2012), usada nesta pesquisa visando entender os assuntos sobre os quais os clientes dos restaurantes discorrem por meio de suas opiniões. Por meio desta técnica, uma série de tópicos foi elencada e, nestes tópicos, são agrupados os termos mais prováveis. Com isso, foi possível confirmar de maneira empírica nesta pesquisa, o que já vinha sendo observado em análises anteriores disposta na literatura sobre modelagem de tópicos: os assuntos sobre os quais as pessoas falavam.

O potencial da Modelagem de Tópicos em revelar a estrutura de tópicos latentes contribui com a intenção desta pesquisa no sentido de revelar conhecimento a partir das opiniões dos clientes, além de ajudar a corroborar os indícios encontrados em várias análises anteriormente realizadas em etapas anteriores à pesquisa executada nesta dissertação.

Também cabe ressaltar o papel das visualizações geradas com o intuito de representar visualmente aspectos importantes do conhecimento dos clientes. Visando extrair conhecimento a partir de coleções de informações, técnicas de visualização permitem construir estruturas mentais ou cognitivas para um domínio específico (ZHU; CHEN, 2005). Em complemento, destaca-se que técnicas de visualização também ajudam a ressaltar características importantes de um conjunto de informações, transformando dados abstratos em formas visuais, dando assim apoio à descoberta de conhecimento em dados (CRAFT; CAIRNS, 2008).

Nesta pesquisa, diversos tipos de visualização foram gerados visando proporcionar uma visão mais aprofundada dos dados. Em alguns casos foram geradas visualizações objetivando uma visão mais ampla da massa de dados e em outros, estas visualizações tinham como propósito revelar as relações existentes entre partes, a exemplo das redes de palavras representadas nesta pesquisa como grafos (OLMEDA-GÓMEZ, 2014).

Por meio da visualização de grafos foi possível encontrar várias conexões entre termos que apareciam em várias das análises realizadas anteriormente, porém nunca dispostas da forma como visualizadas nos grafos. Isto leva a crer que o emprego de bigramas aliado à visualização de grafos é uma excelente forma de revelar relações existentes entre termos de um *corpus*, relações estas que, quando abordadas de outras formas, não proporcionariam a mesma riqueza de informações. Há de se considerar ainda que a facilidade de interpretação das visualizações expostas nesta pesquisa pode ter especial apelo junto aos proprietários de pequenas e médias empresas, conforme o perfil considerado neste estudo.

Esta conclusão corrobora a visão de Olmeda-Gómez (2014), que afirma que a visualização de informações em forma de grafos pode ajudar a revelar relações e padrões existentes entre os elementos de um determinado conjunto de dados, sem que seja necessário qualquer conhecimento prévio sobre estes, o que se dá por meio de estruturas simples, elegantes e diretas.

Por último, destaca-se que, mesmo que os conjuntos de dados trabalhados nesta pesquisa demonstrem certo grau de desequilíbrio na proporção de opiniões positivas em relação às negativas, muitas das descobertas mais interessantes decorrentes das análises ocorreram ao se observar os termos e expressões classificadas como negativos, mesmo estes sendo oriundos de documentos classificados como positivos.

Segundo a literatura, o conhecimento dos clientes, especialmente aqueles que demonstram descontentamentos, pode ser utilizado para melhorar vários aspectos das empresas,

além de constituírem-se numa forma rica e gratuita de obtenção de opiniões e ideias de melhoria de forma gratuita para a empresa (HOROVITZ, 2000; MICHELLI, 2011).

Ao considerar a classificação do conhecimento do cliente sugerida por Gebert *et al.* (2003), conclui-se que esta pesquisa traz contribuições ao campo de estudo do Conhecimento do Cliente e, mais especificamente, em relação ao conhecimento do cliente e conhecimento sobre o cliente.

5 CONCLUSÃO E TRABALHOS FUTUROS

A importância de considerar o conhecimento proveniente do cliente para o desenvolvimento do negócio é indiscutível às empresas de sucesso (DESOUZA; AWAZU, 2005). Cada vez mais, nota-se esforços das empresas no sentido de desenvolver estratégias que possibilitem, ao mesmo tempo, ampliar relacionamentos com os clientes, bem como utilizar suas ideias com o intuito de melhorar produtos e serviços (CHUA; BANERJEE, 2013).

A contribuição dos clientes evidencia-se por meio de vários aspectos que afetam diretamente a performance das empresas (KUMAR *et al.*, 2010), considerando que atualmente produtos e serviços fornecidos pelas empresas estejam cada vez mais parecidos, sendo que a preferência de um cliente por uma ou outra marca se dá por pequenos detalhes (DAVENPORT; HARRIS, 2007), na maior parte das vezes subjetivos e, portanto, relacionados ao arcabouço de conhecimentos tácitos dos clientes (BLACKWELL; MINIARD; ENGEL, 2006).

Assim, alcançar o conhecimento tácito dos clientes é benéfico e essencial (NONAKA, 2006), sobretudo como forma de garantir a sobrevivência das pequenas e médias empresas, dado que este tipo de organização costuma não ter o mesmo nível de recursos que as grandes empresas dispõem (DESOUZA; AWAZU, 2005).

Neste contexto, as ferramentas da Web 2.0, sobretudo as Redes Sociais, cumprem um papel crucial, e as empresas dependem cada vez mais destas para interagir com seus clientes (CHOUDHURY; HARRIGAN, 2014), que por sua vez, adotam cada vez mais as Redes Sociais como forma de buscar informações sobre empresas, produtos ou serviços e publicar opiniões sobre suas experiências com estas.

Entretanto, muito embora a quantidade de opiniões de clientes disponíveis em redes sociais a torne uma fonte valiosa de informações, a tarefa de analisar estes dados não é banal e, considerando-se o volume crescente desses dados, não se trata de tarefa que possa ser desenvolvida de forma manual pelas empresas, notadamente as de pequeno e médio portes.

Neste contexto, apresenta-se o cenário que motivou esta pesquisa de dissertação: conciliar as técnicas de mineração de textos com o objetivo de revelar conhecimento a partir das opiniões dos usuários de rede sociais, sobretudo opiniões relacionadas às suas experiências com restaurantes, domínio escolhido para o desenvolvimento desta pesquisa.

Assim, esta pesquisa apresentou um *framework* para mineração de texto para descoberta de conhecimento do cliente referente às suas experiências em restaurantes oriundas de redes sociais, aplicável à realidade de pequenas e médias empresas. Como principal

resultado, destaca-se a geração de visualizações gráficas que contribuíram para evidenciar relações latentes entre diversas expressões e termos que não eram óbvias e que foram descobertos a partir das análises do *framework* delineado.

A Análise de Sentimentos aliada à Modelagem de Tópicos revelou que os aspectos mais abordados pelos clientes referem-se à comida, ao lugar e o atendimento, variando em intensidade e polaridade. Em complemento, pôde-se inferir conhecimento sobre as inclinações das opiniões dos clientes em diferentes contextos.

A literatura é rica em termos de aplicações para Análise de Sentimentos. Entretanto, a maioria dos trabalhos que apresentam técnicas e ferramentas descrevem aplicações no idioma inglês. Esta pesquisa apresenta uma aplicação de Análise de Sentimentos em língua portuguesa que pode ser considerada bem-sucedida, muito embora haja limitações. Nesta pesquisa a Análise de Sentimentos é usada como forma de inferir conhecimento sobre o que os clientes falam e como se sentem ao falar sobre determinados aspectos, o que corrobora as ideias de Pang e Lee (2008) e Liu (2012), que apregoam que a simplicidade do método não subjugue sua eficácia.

Outro aspecto importante implementado nesta pesquisa tange à aplicação da modelagem de tópicos, por meio da técnica conhecida como *Latent Dirichlet Allocation* (BLEI, 2012), que no contexto desta pesquisa cumpre o papel de revelar o que os clientes falam quando comentam algo relacionado às suas experiências. Os resultados encontrados condizem com a ideia fundamental por trás do modelo probabilístico de Blei (2012), muito embora este tenha sido aplicado de forma objetiva, gerando assim poucos tópicos com apenas alguns itens em cada um. Entretanto, o modelo elaborado cumpre seu desígnio, sendo capaz de indicar nos tópicos abordados, uma série de termos que indicam seu conteúdo, o que, por sua vez, permite inferir o tema do referido tópico em questão.

Os dados utilizados neste experimento foram extraídos diretamente da rede social TripAdvisor Brasil, escolhida para a realização desta pesquisa por conter opiniões de usuários sobre suas experiências em diversos tipos de empreendimentos, dentre os quais se destacam hotéis, restaurantes e empresas prestadoras de serviços. Para a extração dos dados da rede social, a pesquisa realizada fez uso da técnica de *web scraping*, que consiste na extração direta por meio de um script em R, apresentado no Apêndice A, ao final desta dissertação.

O tratamento dos dados realizado nesta pesquisa se deu em duas etapas, tendo sido substancialmente facilitado pela abordagem adotada, que se constitui em si, uma contribuição deste trabalho do ponto de vista técnico em relação à organização dos dados. Estes foram

tratados desde sua extração segundo os princípios do *tidy data*, tendo como consequência direta maior flexibilidade de transformação e tratamento.

Ao se empregar a abordagem *tidy data* para lidar com os dados desta pesquisa, notou-se um ganho de produtividade notável, o que permitiu, entre outras coisas, experimentar diversas formas de realização de determinadas operações de transformação de dados. Tais ganhos de produtividade reforçam sua expressividade como contribuição desta pesquisa, se considerada a afirmação de Dasu e Johnson (2003), que indicam que cerca de 80% do tempo da análise de dados são gastos no processo de limpeza e preparação dos dados brutos. Este, definitivamente não foi o caso deste trabalho, o que sugere que a abordagem *tidy data*, muito embora tenha sido pensada originalmente para lidar com dados estruturados, pode e deva ser melhor explorada em aplicações que envolvam dados não estruturados. Assim, conclui-se que a aplicação da abordagem *tidy data* para dados não estruturados pode contribuir para facilitar diversos processos de mineração de textos, que segundo Feldman e Sanger (2007), são descritos como um trabalho que requer muito esforço por parte dos cientistas de dados.

Quanto às limitações desta pesquisa, destacam-se alguns fatores. O primeiro tange ao universo abordado neste trabalho: Pequenas e Médias Empresas. Muito embora os métodos aplicados nesta pesquisa não se apliquem única e exclusivamente a empresas desse porte, constando-se na literatura aplicações nos mais diversos segmentos e tamanhos de empresas, convém ressaltar que algumas decisões sobre sua construção do *Framework* para Mineração de Opiniões aplicado nesta pesquisa foram influenciadas pelo tamanho das empresas escolhidas. Isto se refletiu, por exemplo, na escolha do método de Análise de Sentimentos, que no contexto desta pesquisa apoiou-se no emprego de léxicos de polaridade para avaliar a inclinação do sentimento das opiniões dos clientes dos restaurantes cujas massas de dados foram analisadas.

Outra limitação desta pesquisa refere-se justamente ao uso de léxicos de polaridade de sentimentos. Considerando que não há muitos léxicos para o idioma português do Brasil, esta pesquisa aplicou os dois mais constantes na literatura: opLexicon versão 3.0 (SOUZA; VIEIRA, 2012) e o sentiLex (SILVA *et al.*, 2010). Ambos revelaram-se limitados, sobretudo ao considerar que nenhum deles foi criado especificamente para o domínio abordado nesta pesquisa, qual seja, restaurantes.

Neste ponto, aborda-se outra limitação desta pesquisa: o domínio no qual foi aplicada (restaurantes). A literatura afirma que há certa constância em relação ao sentimento do cliente em diferentes contextos (PANG; LEE, 2008). Neste aspecto, aponta-se como limitação o uso de léxicos que não foram pensados para o domínio de restaurantes, mas possuem classificações de termos comuns a todo e qualquer contexto.

Em se tratando do conjunto de dados analisado em cada empresa pesquisada, convém lembrar que os mesmos refletem diretamente as entidades analisadas nesta pesquisa, ou seja, cada um dos *corpus* abordados nesta pesquisa possui em seu conteúdo, referências diretas às empresas a partir das quais se extraíram os dados analisados. Por esta razão, convém reforçar que qualquer inferência que extrapole os limites dos objetos estudados nesta pesquisa, pode não ser totalmente precisa para outros objetos, mesmo outros restaurantes. Por último, deve-se lembrar que o universo abordado neste trabalho constitui-se em si numa limitação, considerando que apenas três empresas da cidade de São Paulo foram estudadas.

Os resultados alcançados por esta pesquisa reforçam a importância de se investir em formas de compreender o que os clientes falam sobre o negócio nas redes sociais. Entretanto, somente isto não é suficiente. O cliente pode fornecer conhecimento único que permite com que as empresas aprendam e melhorem suas operações internas (PAQUETTE, 2011). Porém, as empresas devem estar cientes de que não podem tratar seus clientes como indivíduos estáticos, dado que a natureza da relação entre clientes e organizações é dinâmica e, assim como as organizações mudam ao longo do tempo, os clientes também mudam e, por consequência, suas preferências, desejos, estilos de vida, condições e canais de interação (NEJATIAN *et al.*, 2011).

Assume-se que a contribuição deste trabalho para o campo de estudo do Conhecimento do Cliente resida na aplicação prática da Mineração de Textos para revelar padrões e ainda possibilitar a descoberta de conhecimento a partir das opiniões de clientes extraídas de redes sociais.

Destaca-se a aplicação do *framework* apresentado nesta pesquisa como contribuição Acadêmica para o domínio de restaurantes, principalmente Pequenos e Médios. Sobretudo considerando que, em geral, os gerentes e proprietários de estabelecimentos deste tipo e segmento desconhecem os benefícios das ferramentas e técnicas demonstradas nesta dissertação.

O *framework* empregado provou-se útil como ferramenta para compreender melhor o cliente, suas expectativas e até mesmo suas frustrações, possibilitando obter conhecimento acerca dos clientes para benefício das empresas.

Como sugestão de trabalhos futuros propõe-se a expansão do *framework* e sua implementação em uma linguagem, como o Python, dado que esta é extremamente popular e possui centenas de implementações de soluções para lidar com aprendizado de máquina e outros algoritmos sofisticados e com melhor performance para desenvolvimento de sistemas.

Em se tratando de inteligência computacional, seria uma interessante adição o emprego de certas técnicas inteligentes, sobretudo visando o desenvolvimento de um sistema mais robusto e que apreenda com os dados, com o objetivo de criar um sistema de tomada de decisões de comunicação para Pequenas e Médias Empresas.

Pesquisas futuras poderiam ainda avaliar em especial as técnicas de Análise de Agrupamentos aliadas à mineração de opiniões. Outro fator que poderia gerar boas perspectivas de pesquisa tange à aplicação de *stemmer*, que nesta pesquisa foi descartado por razões já explicitadas, muito embora venha a ser uma técnica importante se considerada uma massa de dados exponencialmente maior.

Outra sugestão volta-se à aplicação deste sistema em diferentes domínios. Trata-se de um desafio, dado que, como já comentado, o domínio tem grande influência sobre os resultados alcançados. Entretanto, acredita-se que ao aplicar-se o conjunto correto de técnicas, sobretudo envolvendo aprendizado de máquina, os resultados possam ser satisfatórios.

REFERÊNCIAS

ABBASI, A. et al. Selecting Attributes for Sentiment Classification Using Feature Relation Networks. **IEEE Transactions on Knowledge and Data Engineering**, v. 23, n. 3, p. 447–462, mar. 2011.

ABBASI, A.; CHEN, H.; SALEM, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. **ACM Transactions on Information Systems**, v. 26, n. 3, p. 1–34, 1 jun. 2008.

AGGARWAL, C. C.; ZHAI, C. (EDS.). **Mining Text Data**. Boston, MA: Springer US, 2012.

ALBA, J. W.; HUTCHINSON, J. W. Dimensions of consumer expertise. **Journal of consumer research**, v. 13, n. 4, p. 411–454, 1987.

ALBA, J. W.; HUTCHINSON, J. W. Knowledge Calibration: What Consumers Know and What They Think They Know. **Journal of Consumer Research**, v. 27, n. 2, p. 123, 2000.

ALJALIDI, N.; ALSHEDOKHI, M.; SABA, T. The impact of tripadvisor ratings for tourisim advertisement in Saudi Arabia. **J Bus Technov**, v. 4, p. 90–95, 2016.

ARANHA, C. N. **Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional**. [s.l.] PUC-Rio, 2007.

BAARS, H.; KEMPER, H.-G. Management Support with Structured and Unstructured Data—An Integrated Business Intelligence Framework. **Information Systems Management**, v. 25, n. 2, p. 132–148, 28 mar. 2008.

BACKSTROM, L. et al. **Four degrees of separation**. Proceedings of the 4th Annual ACM Web Science Conference. **Anais**...ACM, 2012

BAUMGARTNER, R. et al. **Web data extraction for business intelligence: the lixto approach**. In Proc. of BTW 2005. **Anais**...2005

BAUMGARTNER, R.; GATTERBAUER, W.; GOTTLOB, G. Web data extraction system. In: **Encyclopedia of Database Systems**. [s.l.] Springer, 2009. p. 3465–3471.

BERRY, M. W.; KOGAN, J. (EDS.). **Text mining: applications and theory**. Chichester, U.K: Wiley, 2010.

BHATTACHARYA, C. B.; SEN, S. Consumer-Company Identification: A Framework for Understanding Consumers' Relationships with Companies. **Journal of Marketing**, v. 67, n. 2, p. 76–88, 2003.

BLACKWELL, R. D.; MINIARD, P. W.; ENGEL, J. F. **Consumer Behavior**. [s.l.] Thomson/South-Western, 2006.

BLANCHARD, O. **Social media ROI: managing and measuring social media efforts in your organization**. Indianapolis, Ind: Que Publ, 2011.

BLEI, D. M. Probabilistic topic models. **Communications of the ACM**, v. 55, n. 4, p. 77, 1 abr. 2012.

BLEI, D. M.; LAFFERTY, J. D. Topic models. **Text mining: classification, clustering, and applications**, v. 10, n. 71, p. 34, 2009.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

BOYD, D. **Taken Out of Context: American Teen Sociality in Networked Publics**. Rochester, NY: Social Science Research Network, 9 dez. 2008. Disponível em: <http://dx.doi.org/10.2139/ssrn.1344756>.

BOYD, D. Social Network Sites as Networked Publics: Affordances, Dynamics, and Implications. In: **Networked Self: Identity, Community, and Culture on Social Network Sites**. [s.l.] Zizi Papacharissi, 2010. p. 39–58.

BRADLEY, A. The Six Core Principles of Social-Media-Based Collaboration, G00172930. **Gartner Inc., Stamford, CT**, 2009.

BRUCKS, M. The Effects of Product Class Knowledge on Information Search Behavior. **Journal of Consumer Research**, v. 12, 1 jun. 1985.

CAMPBELL, A. J. Creating customer knowledge competence: managing customer relationship management programs strategically. **Industrial marketing management**, v. 32, n. 5, p. 375–383, 2003.

CASEY, C.; LI, J. Web 2.0 Technologies and Authentic Public Participation: Engaging Citizens in Decision Making Processes. **http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/978-1-4666-0318-9.ch011**, p. 197–223, 2012.

CASEY, S. **The 2016 Nielsen social media report**. [s.l.] The Nielsen Company, 2017. Disponível em: <http://www.nielsen.com/us/en/insights/reports/2017/2016-nielsen-social-media-report.html>.

CASTELLANOS, M. et al. **LCI: a social channel analysis platform for live customer intelligence**. Proceedings of the 2011 ACM SIGMOD International Conference on Management of data. **Anais**...ACM, 2011

CATANESE, S. A. et al. Crawling Facebook for Social Network Analysis Purposes. **arXiv:1105.6307 [physics]**, p. 1, 2011.

CHEN, H.; CHAU, M.; ZENG, D. CI Spider: a tool for competitive intelligence on the Web. **Decision Support Systems**, v. 34, n. 1, p. 1–17, 2002.

CHIAVETTA, F.; LO BOSCO, G.; PILATO, G. **A Lexicon-based Approach for Sentiment Classification of Amazon Books Reviews in Italian Language:** SCITEPRESS - Science and and Technology Publications, 2016Disponível em: <http://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0005915301590170>. Acesso em: 27 ago. 2017

CHIKERSAL, P.; PORIA, S.; CAMBRIA, E. **SeNTU: Sentiment Analysis of Tweets by Combining a Rule-based Classifier with Supervised Learning.** SemEval@ NAACL-HLT. **Anais**...2015Disponível em: <https://www.aclweb.org/anthology/S/S15/S15-2.pdf#page=689>. Acesso em: 28 ago. 2017

CHOMSKY, N. **Linguagem e mente: pensamentos atuais sobre antigos problemas**. [s.l.] Ed. UNESP, 2009.

CHOUDHURY, M. M.; HARRIGAN, P. CRM to social CRM: the integration of new technologies into customer relationship management. **Journal of Strategic Marketing**, v. 22, n. 2, p. 149–176, 23 fev. 2014.

CHOWDHURY, G. G. Natural language processing. **Annual Review of Information Science and Technology**, v. 37, n. 1, p. 51–89, 2003.

CHUA, A. Y. .; BANERJEE, S. Customer knowledge management via social media: the case of Starbucks. **Journal of Knowledge Management**, v. 17, n. 2, p. 237–249, 29 mar. 2013.

CLARK, A.; FOX, C.; LAPPIN, S. (EDS.). **The handbook of computational linguistics and natural language processing**. Chichester, West Sussex ; Malden, MA: Wiley-Blackwell, 2010.

CODD, E. F. **The relational model for database management: version 2**. Reading, Mass: Addison-Wesley, 1990.

CORMODE, G.; KRISHNAMURTHY, B. Key differences between Web 1.0 and Web 2.0. **First Monday**, v. 13, n. 6, 25 abr. 2008.

CORREA, T.; HINSLEY, A. W.; ZÚÑIGA, H. G. DE. Who interacts on the Web?: The intersection of users' personality and social media use. **Computers in Human Behavior**, v. 26, n. 2, p. 247–253, 2010.

CRAFT, B.; CAIRNS, P. **Directions for methodological research in information visualization**. Information Visualisation, 2008. IV'08. 12th International Conference. **Anais**...IEEE, 2008

CROWDFLOWER. **2016 Data Science Report**. [s.l.] CrowdFlower, 2016.

CSARDI, G.; NEPUSZ, T. The igraph software package for complex network research. **InterJournal**, v. Complex Systems, p. 1695, 2006.

DALKIR, K.; LIEBOWITZ, J. **Knowledge Management in Theory and Practice**. second edition ed. Cambridge, Mass: The MIT Press, 2011.

DARROCH, J.; MCNAUGHTON, R. Beyond market orientation: Knowledge management and the innovativeness of New Zealand firms. **European Journal of Marketing**, v. 37, n. 3/4, p. 572–593, abr. 2003.

DASU, T.; JOHNSON, T. **Exploratory Data Mining and Data Cleaning**. [s.l.] Wiley, 2003.

DAVENPORT, T. H.; HARRIS, J. G. **Competing on Analytics: The New Science of Winning**. 1 edition ed. Boston, Mass: Harvard Business Review Press, 2007.

DAVENPORT, T. H.; PRUSAK, L. **Working Knowledge: How Organizations Manage What They Know**. 2nd edition ed. Boston, Mass: Harvard Business Review Press, 2000.

DE LONG, D. W.; FAHEY, L. Diagnosing cultural barriers to knowledge management. **The Academy of management executive**, v. 14, n. 4, p. 113–127, 2000.

DEERWESTER, S. et al. Indexing by latent semantic analysis. **Journal of the American society for information science**, v. 41, n. 6, p. 391, 1990.

DESOUZA, K. C.; AWAZU, Y. What do they know? **Business strategy review**, v. 16, n. 1, p. 41–45, 2005.

DESOUZA, K. C.; AWAZU, Y. Knowledge management at SMEs: five peculiarities. **Journal of Knowledge Management**, v. 10, n. 1, p. 32–43, jan. 2006.

DESSART, L.; VELOUTSOU, C.; MORGAN-THOMAS, A. Consumer engagement in online brand communities: a social media perspective. **Journal of Product & Brand Management**, v. 24, n. 1, p. 28–42, 16 mar. 2015.

DEVIKA, K.; SURENDRAN, S. An overview of web data extraction techniques. **International journal of scientific engineering and technology**, v. 2, n. 4, 2013.

DING, X.; LIU, B.; YU, P. S. **A holistic lexicon-based approach to opinion mining**. Proceedings of the 2008 international conference on web search and data mining. **Anais**...ACM, 2008Disponível em: <http://dl.acm.org/citation.cfm?id=1341561>. Acesso em: 27 ago. 2017

DRUCKER, P. F. **The Age of Discontinuity: Guidelines to Our Changing Society**. 2nd Revised ed. ed. [s.l.] Transaction Publishers, 1992.

DURST, S.; WILHELM, S. Do you know your knowledge at risk? **Measuring Business Excellence**, v. 17, n. 3, p. 28–39, 23 ago. 2013.

ENGINKAYA, E.; YILMAZ, H. What Drives Consumers to Interact with Brands through Social Media? A Motivation Scale Development Study. **Procedia - Social and Behavioral Sciences**, v. 148, p. 219–226, ago. 2014.

FAHRNI, A.; KLENNER, M. **Old wine or warm beer: Target-specific sentiment analysis of adjectives**. Proc. of the Symposium on Affective Language in Human and Machine, AISB. **Anais**...2008Disponível em: <http://aisb.org.uk/convention/aisb08/proc/proceedings/02%20Affective%20Language/Final%20vol%2002.pdf#page=66>. Acesso em: 28 ago. 2017

FALBEL, D. **ptstem: Stemming Algorithms for the Portuguese Language**. [s.l: s.n.].

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

FEINERER, I.; HORNIK, K. **tm: Text mining package**. [s.l: s.n.].

FELDMAN, R.; SANGER, J. **The text mining handbook: advanced approaches in analyzing unstructured data**. Cambridge; New York: Cambridge University Press, 2007.

FELLOWS, I. **wordcloud: Word Clouds**. [s.l: s.n.].

FERNANDES, F. A. R. A indústria hoteleira e as reclamações online: o caso do TripAdvisor. 2015.

GARCÍA-MURILLO, M.; ANNABI, H. Customer Knowledge Management. **The Journal of the Operational Research Society**, v. 53, n. 8, p. 875–884, ago. 2002.

GARDNER, H. **Frames of mind: the theory of multiple intelligences**. New York: Basic Books, 2011.

GARFINKEL, S.; COX, D. **Finding and archiving the internet footprint**. [s.l.] Naval Postgraduate School Monterey CA, 2009.

GATICA-PEREZ, D.; RUIZ-CORREA, S.; SANTANI, D. **What TripAdvisor Can" t Tell: Crowdsourcing Urban Impressions for Whole Cities**. [s.l.] Digital Polis, 2016.

GAUTAM, G.; YADAV, D. **Sentiment analysis of twitter data using machine learning approaches and semantic analysis**. Contemporary computing (IC3), 2014 seventh international conference on. **Anais**...IEEE, 2014Disponível em: <http://ieeexplore.ieee.org/abstract/document/6897213/>. Acesso em: 28 ago. 2017

GEBERT, H. et al. Knowledge-enabled customer relationship management: integrating customer relationship management and knowledge management concepts[1]. **Journal of Knowledge Management**, v. 7, n. 5, p. 107–123, dez. 2003.

GIBSON, J. J. **The Ecological Approach to Visual Perception**. [s.l.] Taylor & Francis Group, 1986.

GIL, A. C. **Métodos e técnicas de pesquisa social**. São Paulo: Atlas, 2008a.

GIL, A. C. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2008b.

GJOKA, M. et al. **Walking in Facebook: A Case Study of Unbiased Sampling of OSNs**. [s.l: s.n.].

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. **CS224N Project Report, Stanford**, v. 1, n. 2009, p. 12, 2009.

GORDON, J.; PATTERSON, J. A. Response to Tracy's Under the "Big Tent": Establishing Universal Criteria for Evaluating Qualitative Research. **Qualitative Inquiry**, v. 19, n. 9, p. 689–695, nov. 2013.

GREENO, J. G. Gibson's affordances. **Psychological review**, v. 101, n. 2, p. 336–342, 1994.

GRISHMAN, R. **Computational Linguistics: An Introduction**. [s.l.] Cambridge University Press, 1986.

GRÜN, B.; HORNIK, K. topicmodels: An R Package for Fitting Topic Models. **Journal of Statistical Software**, v. 40, n. 13, p. 1–30, 2011.

HAN, J.; KAMBER, M. **Data mining: concepts and techniques**. 3rd ed ed. Burlington, MA: Elsevier, 2011.

HATZIVASSILOGLOU, V.; MCKEOWN, K. R. **Predicting the semantic orientation of adjectives**. Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. **Anais**...Association for Computational Linguistics, 1997Disponível em: <http://dl.acm.org/citation.cfm?id=979640>. Acesso em: 27 ago. 2017

HEIMBACH, A. E.; JOHANSSON, J. K.; MACLACHLAN, D. L. Product familiarity, information processing, and country-of-origin cues. **NA-Advances in Consumer Research Volume 16**, 1989.

HOFFMAN, D. L.; FODOR, M. Can you measure the ROI of your social media marketing? **MIT Sloan Management Review**, v. 52, n. 1, p. 41, 2010.

HOFMANN, M.; CHISHOLM, A. **Text Mining and Visualization - Case Studies Using Open-Source Tools**. Boca Raton, FL: Taylor & Francis Group, 2013.

HOFMANN, T. **Probabilistic latent semantic analysis**. Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. **Anais**...Morgan Kaufmann Publishers Inc., 1999Disponível em: <http://dl.acm.org/citation.cfm?id=2073829>. Acesso em: 13 jun. 2017

HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. **Machine learning**, v. 42, n. 1, p. 177–196, 2001.

HOLLEBEEK, L. D. Demystifying customer brand engagement: Exploring the loyalty nexus. **Journal of Marketing Management**, v. 27, n. 7–8, p. 785–807, jul. 2011.

HOLLEBEEK, L. D.; GLYNN, M. S.; BRODIE, R. J. Consumer Brand Engagement in Social Media: Conceptualization, Scale Development and Validation. **Journal of Interactive Marketing**, v. 28, n. 2, p. 149–165, maio 2014.

HOLZINGER, A. et al. Combining HCI, Natural Language Processing, and Knowledge Discovery-Potential of IBM Content Analytics as an assistive technology in the biomedical field. In: **Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data**. [s.l.] Springer, 2013. p. 13–24.

HOROVITZ, J. Using information to bond with customers. In: **Competing with Information: A Manager's Guide to Creating Business Value with Information Content**. Chichester, England: John Wiley & Sons, 2000. p. 35–53.

HU, M.; LIU, B. **Mining and summarizing customer reviews**. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. **Anais**...ACM, 2004Disponível em: <http://dl.acm.org/citation.cfm?id=1014073>. Acesso em: 27 ago. 2017

ISO. **ISO 8601:2004 - Data elements and interchange formats - Information interchange - Representation of dates and times**, 2004. Disponível em: <https://www.iso.org/standard/40874.html>. Acesso em: 21 out. 2017

JACKSON, P.; MOULINIER, I. **Natural Language Processing for Online Applications: Text retrieval, extraction and categorization. Second revised edition**. [s.l.] John Benjamins Publishing Company, 2007.

JAYACHANDRAN, S. et al. The role of relational information processes and technology use in customer relationship management. **Journal of marketing**, v. 69, n. 4, p. 177–192, 2005.

JOHN, N. A. Sharing and Web 2.0: The emergence of a keyword. **new media & society**, v. 15, n. 2, p. 167–182, 2013.

KALETKA, C.; PELKA, B. Web 2.0 revisited: user-generated content as a social innovation. **International Journal of Innovation and Sustainable Development**, v. 5, n. 2–3, p. 264–275, 2011.

KAMANWAR, N. V.; KALE, S. G. **Web data extraction techniques: A review**. Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave), World Conference on. **Anais**...IEEE, 2016

KAPLAN, A. M.; HAENLEIN, M. Users of the world, unite! The challenges and opportunities of Social Media. **Business Horizons**, v. 53, n. 1, p. 59–68, 2010.

KEATES, N. Deconstructing tripadvisor. **Wall Street Journal**, v. 1, n. 4, 2007.

KIETZMANN, J. H. et al. Social media? Get serious! Understanding the functional building blocks of social media. **Business Horizons**, v. 54, n. 3, p. 241–251, 2011.

KIM, S.-M.; HOVY, E. **Determining the sentiment of opinions**. Proceedings of the 20th international conference on Computational Linguistics. **Anais**...Association for Computational Linguistics, 2004Disponível em: <http://dl.acm.org/citation.cfm?id=1220555>. Acesso em: 27 ago. 2017

KLEINBERG, J. **The Small-world Phenomenon: An Algorithmic Perspective**. Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing. **Anais**...: STOC '00.New York, NY, USA: ACM, 2000Disponível em: <http://doi.acm.org/10.1145/335305.335325>

KRUG, S. **Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability**. [s.l.] Pearson Education, 2013.

KUMAR, V. et al. Undervalued or Overvalued Customers: Capturing Total Customer Engagement Value. **Journal of Service Research**, v. 13, n. 3, p. 297–310, 1 ago. 2010.

LAKATOS, E. M.; MARCONI, M. DE A. **Fundamentos de metodologia científica**. São Paulo: Atlas, 2003.

LEONARDI, P. M. Social Media , Knowledge Sharing , and Innovation : Toward a Theory of Communication Visibility. **Information systems Research**, v. 25, n. 4, p. 796–816, 2014.

LEONARDI, P. M. Ambient Awareness and Knowledge Acquisition: Using Social Media to Learn "Who Knows What" and "Who Knows Whom". v. 39, n. 4, p. 747–762, 2015.

LEVY, M. WEB 2.0 implications on knowledge management. **Journal of Knowledge Management**, v. 13, n. 1, p. 120–134, 20 fev. 2009.

LIDDY, E. D. Natural language processing. 2001.

LIN, H.-F. Contextual factors affecting knowledge management diffusion in SMEs. **Industrial Management & Data Systems**, v. 114, n. 9, p. 1415–1437, 7 out. 2014.

LIN, K. Y.; LU, H. P. Why people use social networking sites: An empirical study integrating network externalities and motivation theory. **Computers in Human Behavior**, v. 27, n. 3, p. 1152–1161, 2011.

LIU, B. Sentiment Analysis and Subjectivity. **Handbook of natural language processing**, v. 2, p. 627–666, 2010.

LIU, B. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, v. 5, n. 1, p. 1–167, 2012.

MALIK, S. K.; RIZVI, S. **Information Extraction Using Web Usage Mining, Web Scrapping and Semantic Annotation**. IEEE, out. 2011Disponível em: <http://ieeexplore.ieee.org/document/6112910/>. Acesso em: 6 nov. 2017

MARRES, N.; WELTEVREDE, E. SCRAPING THE SOCIAL?: Issues in live social research. **Journal of Cultural Economy**, v. 6, n. 3, p. 313–335, ago. 2013.

MARTINS, G. DE A.; THEÓPHILO, C. R. **Metodologia da Investigação Científica Para Ciências Sociais Aplicadas**. 3ª Edição ed. Brasil: Atlas, 2017.

MARWICK, A. E. Instafame: Luxury Selfies in the Attention Economy. **Public Culture**, v. 27, n. 1 75, p. 137–160, jan. 2015.

MARWICK, A. E.; BOYD, D. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. **New Media & Society**, v. 13, n. 1, p. 114–133, fev. 2011.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, v. 5, n. 4, p. 1093–1113, dez. 2014.

MICHAELIDOU, N.; SIAMAGKA, N. T.; CHRISTODOULIDES, G. Usage, barriers and measurement of social media marketing: An exploratory investigation of small and medium B2B brands. **Industrial Marketing Management**, v. 40, n. 7, p. 1153–1159, 2011.

MICHELLI, J. A. **The Starbucks experience: 5 principles for turning ordinary into extraordinary**. [s.l.] McGraw Hill Professional, 2011.

MIGUÉNS, J.; BAGGIO, R.; COSTA, C. Social media and tourism destinations: TripAdvisor case study. **Advances in tourism research**, v. 26, n. 28, p. 1–6, 2008.

MILLER, G. A. WordNet: A Lexical Database for English. **Commun. ACM**, v. 38, n. 11, p. 39–41, nov. 1995.

MITCHELL, T. M. **Machine Learning**. New York: McGraw-Hill, 1997.

MUNZERT, S. **Automated data collection with R: a practical guide to Web scraping and text mining**. Chichester, West Sussex, United Kingdom: John Wiley & Sons Inc, 2015.

MUSTO, C.; SEMERARO, G.; POLIGNANO, M. A comparison of lexicon-based approaches for sentiment analysis of microblog posts. **Information Filtering and Retrieval**, v. 59, 2014.

NEJATIAN, H. et al. The Influence of Customer Knowledge on CRM Performance of Malaysian ICT Companies: A Structural Equation Modeling Approach. **International Journal of Business and Management**, v. 6, n. 7, 30 jun. 2011.

NEVES, P. I.; CORRÊA, D. A.; CAVALCANTI, M. C. Uma análise sobre abordagens e ferramentas para Extração de Informação. **Revista Militar de Ciência e Tecnologia**, n. 3º Trim, p. 32–58, 2013.

NEWMAN, M. E. The structure and function of complex networks. **SIAM review**, v. 45, n. 2, p. 167–256, 2003.

NIELSEN, J.; LORANGER, H. **Usabilidade na web**. [s.l.] CAMPUS - RJ, 2007.

NONAKA, I. A Dynamic Theory of Organizational Knowledge Creation. **Organization Science**, v. 5, n. 1, p. 14–37, fev. 1994.

NONAKA, I. Organizational Knowledge Creation Theory: Evolutionary Paths and Future Advances. **Organization Studies**, v. 27, n. 8, p. 1179–1208, 6 jun. 2006.

NONAKA, I.; TAKEUCHI, H. **The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation**. 1. ed. [s.l.] Oxford University Press, 1995.

NORMAN, D. A. **The Design of Everyday Things**. [s.l: s.n.]. v. 16

NUNES, M. B. et al. Knowledge management issues in knowledge-intensive SMEs. **Journal of Documentation**, v. 62, n. 1, p. 101–119, jan. 2006.

O'CONNOR, P. User-generated content and travel: A case study on Tripadvisor. com. **Information and communication technologies in tourism 2008**, p. 47–58, 2008.

OLMEDA-GÓMEZ, C. Visualización de información. **El profesional de la información**, v. 23, n. 3, 2014.

O'REILLY, T. **What Is Web 2.0**. Disponível em: <http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html>. Acesso em: 5 set. 2015.

O'REILLY, T. What is Web 2.0: Design patterns and business models for the next generation of software. **Communications & strategies**, n. 1, p. 17, 2007.

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD). **SME statistics: towards a more systematic statistical measurement of SME behavior**. Istanbul, Turkey: 2nd OECD Conference of Ministers Responsible for Small and Medium Enterprises (SMEs), 2004. Disponível em: <http://www.oecd.org/cfe/smes/31919286.pdf>.

PANG, B.; LEE, L. **A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts**. Proceedings of the 42nd annual meeting on Association for Computational Linguistics. **Anais**...Association for Computational Linguistics, 2004Disponível em: <http://dl.acm.org/citation.cfm?id=1218990>. Acesso em: 11 jun. 2017

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. **Found. Trends Inf. Retr.**, v. 2, n. 1–2, p. 1–135, jan. 2008.

PANG, B.; LEE, L.; VAITHYANATHAN, S. **Thumbs up?: sentiment classification using machine learning techniques**. Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. **Anais**...Association for Computational Linguistics, 2002Disponível em: <http://dl.acm.org/citation.cfm?id=1118704>. Acesso em: 27 ago. 2017

PAQUETTE, S. Customer Knowledge Management. In: **Encyclopedia of knowledge management**. Hershey, PA: Idea Group Reference, 2011.

PARK, S.; KIM, Y. **Building thesaurus lexicon using dictionary-based approach for sentiment classification**. Software Engineering Research, Management and Applications (SERA), 2016 IEEE 14th International Conference on. **Anais**...IEEE, 2016Disponível em: <http://ieeexplore.ieee.org/abstract/document/7516126/>. Acesso em: 27 ago. 2017

PEDERSEN, T. L. **ggraph: An Implementation of Grammar of Graphics for Graphs and Networks**. [s.l: s.n.].

PINKER, S. **Como a mente funciona**. [s.l.] Companhia das Letras, 1998.

PINKER, S. **O Instinto Da Linguagem: Como a Mente Cria a Linguagem**. [s.l.] Martins Fontes, 2002.

PINKER, S. **Do que é feito o pensamento: a língua como janela para a natureza humana**. [s.l.] COMPANHIA DAS LETRAS, 2008.

PLAKE, C. et al. AliBaba: PubMed as a graph. **Bioinformatics**, v. 22, n. 19, p. 2444–2445, 2006.

POLANYI, M. **The Tacit Dimension**. [s.l.] University of Chicago Press, 1966.

PONGPAEW, W.; SPEECE, M.; TIANGSOONGNERN, L. Social presence and customer brand engagement on Facebook brand pages. **Journal of Product & Brand Management**, v. 26, n. 3, p. 262–281, 15 maio 2017.

PRAHALAD, C. K.; RAMASWAMY, V. Co-creating unique value with customers. **Strategy & leadership**, v. 32, n. 3, p. 4–9, 2004.

PRODANOV, C. C.; FREITAS, E. C. DE. **Metodologia do Trabalho Científico: Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico-2ª Edição**. [s.l.] Editora Feevale, 2013.

PYLE, D. **Data preparation for data mining**. [s.l.] morgan kaufmann, 1999. v. 1

READ, J.; CARROLL, J. **Weakly supervised techniques for domain-independent sentiment classification**. Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. **Anais**...ACM, 2009Disponível em: <http://dl.acm.org/citation.cfm?id=1651470>. Acesso em: 28 ago. 2017

ROBINSON, D. **The Impressive Growth of R**. Disponível em: <https://stackoverflow.blog/2017/10/10/impressive-growth-r/>. Acesso em: 21 out. 2017.

ROLLINS, M.; HALINEN, A. **Customer knowledge management competence: Towards a theoretical framework**. Proceedings of the 38th Annual Hawaii International Conference on System Sciences. **Anais**...IEEE, 2005Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1385729>. Acesso em: 15 ago. 2016

ROWLEY, J. Eight questions for customer knowledge management in e-business. **Journal of Knowledge Management**, v. 6, n. 5, p. 500–511, dez. 2002.

RSTUDIO TEAM. **RStudio: Integrated Development Environment for R**. Boston, MA: RStudio, Inc., 2015.

SALOMANN, H. et al. Rejuvenating Customer Management: How to Make Knowledge For, From and About Customers Work. **European Management Journal**, v. 23, n. 4, p. 392–403, ago. 2005.

SCHIVINSKI, B.; DABROWSKI, D. The effect of social media communication on consumer perceptions of brands. **Journal of Marketing Communications**, v. 22, n. 2, p. 189–214, 3 mar. 2016.

SEBRAE. **Participação das micro e pequenas empresas**. Disponível em: <http://www.sebrae.com.br/>. Acesso em: 20 set. 2016.

SEBRAE-SP. **Panorama dos Pequenos Negocios 2017**. [s.l.] SEBRAE, 2017.

SILGE, J.; ROBINSON, D. **Text Mining with R: A Tidy Approach**. [s.l.] O'Reilly Media, Incorporated, 2017.

SILVA, E. L. DA; MENEZES, E. M. Metodologia da pesquisa e elaboração de dissertação. **Florianópolis, UFSC**, v. 5, n. 6, 2005.

SILVA, L. A. DA; PERES, S. M.; BOSCARIOLI, C. **Introdução à Mineração de Dados: Com Aplicações em R**. [s.l.] Elsevier Brasil, 2017.

SILVA, M. J. et al. **Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis**. [s.l.] University of Lisbon, Faculty of Sciences, LASIGE, dez. 2010. Disponível em: <http://hdl.handle.net/10455/6694>.

SINGH, J.; GUPTA, V. A systematic review of text stemming techniques. **Artificial Intelligence Review**, v. 48, n. 2, p. 157–217, ago. 2017.

SMITS, M.; MOGOS, S. The Impact of Social Media on Business Performance. **European Conference on Information Systems (2013)**, p. 1–12, 2014.

SMYTH, P. C. B.; WU, G.; GREENE, D. Does TripAdvisor makes hotels better. **Derek Greene School of Computer Science & Informatics, University College Dublin Belfield**, 2010.

SOUZA, M.; VIEIRA, R. Sentiment Analysis on Twitter Data for Portuguese Language. In: CASELI, H. et al. (Eds.). . **Computational Processing of the Portuguese Language**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. v. 7243p. 241–247.

STELZNER, M. **2016 social media marketing industry report - How Marketers Are Using Social Media to Grow Their Businesses**. [s.l: s.n.]. Disponível em:

<https://pdfs.semanticscholar.org/c900/ae47192dc2d2527520474237ab53ee22c3ed.pdf>. Acesso em: 21 mar. 2017.

STEWART, T. A. **Intellectual Capital: The New Wealth of Organizations**. [s.l.] Doubleday / Currency, 1997.

STRAUSS, A.; CORBIN, J. M. **Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory**. [s.l.] SAGE Publications, 1998.

TABOADA, M. et al. Lexicon-based methods for sentiment analysis. **Computational linguistics**, v. 37, n. 2, p. 267–307, 2011.

TAKEUCHI, H.; NONAKA, I. **Gestão do conhecimento**. [s.l.] Bookman Editora, 2009.

TANG, B.; KAY, S.; HE, H. Toward Optimal Feature Selection in Naive Bayes for Text Categorization. **IEEE Transactions on Knowledge and Data Engineering**, v. 28, n. 9, p. 2508–2521, 1 set. 2016.

THAKKAR, H.; PATEL, D. Approaches for sentiment analysis on twitter: A state-of-art study. **arXiv preprint arXiv:1512.01043**, 2015.

TREEM, J. W.; LEONARDI, P. M. Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association. **Communication Yearbook**, n. 36, p. 143–189, 2012.

TRIPADVISOR. **TripAdvisor Brasil**. Rede Social. Disponível em: <http://www.tripadvisor.com.br/>. Acesso em: 15 ago. 2017.

TURNEY, P. D. **Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews**. Proceedings of the 40th annual meeting on association for computational linguistics. **Anais**...Association for Computational Linguistics, 2002Disponível em: <http://dl.acm.org/citation.cfm?id=1073153>. Acesso em: 27 ago. 2017

VAN NES, F. et al. Language differences in qualitative research: is meaning lost in translation? **European Journal of Ageing**, v. 7, n. 4, p. 313–316, dez. 2010.

VARGIU, E.; URRU, M. Exploiting web scraping in a collaborative filtering- based approach to web advertising. **Artificial Intelligence Research**, v. 2, n. 1, 20 nov. 2012.

VENDEMIA, M. A. When do consumers buy the company? Perceptions of interactivity in company-consumer interactions on social networking sites. **Computers in Human Behavior**, v. 71, p. 99–109, jun. 2017.

VITAK, J. The Impact of Context Collapse and Privacy on Social Network Site Disclosures. **Journal of Broadcasting & Electronic Media**, v. 56, n. 4, p. 451–470, out. 2012.

WAGNER, D.; VOLLMAR, G.; WAGNER, H.-T. The impact of information technology on knowledge creation: An affordance approach to social media. **Journal of Enterprise Information Management**, v. 27, n. 1, p. 31–44, 2014.

WANG, C.; LEE, M. K. O.; HUA, Z. A theory of social media dependence: Evidence from microblog users. **Decision Support Systems**, v. 69, n. JANUARY 2015, p. 40–49, 2015.

WEISS, S. M.; INDURKHYA, N.; ZHANG, T. **Fundamentals of Predictive Text Mining**. London: Springer London, 2010.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. [s.l.] Springer-Verlag New York, 2009.

WICKHAM, H. Tidy data. **Journal of Statistical Software**, v. 59, n. 10, p. 1–23, 2014.

WICKHAM, H. **rvest: Easily Harvest (Scrape) Web Pages**. [s.l: s.n.].

WICKHAM, H. **tidyverse: Easily Install and Load "Tidyverse" Packages**. [s.l: s.n.].

WICKHAM, H.; GROLEMUND, G. **R for Data Science: Import, Tidy, Transform, Visualize, and Model Data**. [s.l.] O'Reilly Media, Inc., 2016.

WIEDEMANN, G. **Text Mining for Qualitative Data Analysis in the Social Sciences**. Wiesbaden: Springer Fachmedien Wiesbaden, 2016.

ZANASI, A. Competitive intelligence through data mining public sources. **Competitive intelligence review**, v. 9, n. 1, p. 44–54, 1998.

ZEMBIK, M. Social media as a source of knowledge for customers and enterprises. **Online Journal of Applied Knowledge Management**, v. 2, n. 2, p. 132–148, 2014.

ZHAI, C.; MASSUNG, S. **Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining**. New York, NY, USA: Association for Computing Machinery and Morgan & Claypool, 2016.

ZHANG, S.; ZHANG, C.; YANG, Q. Data preparation for data mining. **Applied Artificial Intelligence**, v. 17, n. 5–6, p. 375–381, 2003.

ZHU, B.; CHEN, H. Information visualization. **Annual review of information science and technology**, v. 39, n. 1, p. 139–177, 2005.

APÊNDICES

APÊNDICE A – *Script de extração das opiniões de usuários do site TripAdvisor Brasil*

O Apêndice A apresenta o script utilizado nesta pesquisa para extrair as opiniões de usuários da rede social TripAdvisor Brasil. O script foi desenvolvido na linguagem R e as opiniões extraídas com o auxílio do script foram submetidas posteriormente aos processos de pré-processamento e mineração de textos por meio do *framework* de mineração de opiniões apresentado nesta pesquisa.

APÊNDICE B – *Lista final de stop words aplicada na pesquisa*

O Apêndice B apresenta a lista *de stop words* adotada na fase de pré-processamento de todos os dados tratados nesta pesquisa. A lista é proveniente da do pacote **tm**, criado por Ingo Feinerer e Kurt Hornik (2017). A lista teve apenas duas adições: os termos ‘q’ e ‘vc’, dado que os mesmos eram muito constantes e irrelevantes para a obtenção de conhecimento útil da massa de dados analisada.

APÊNDICE A – Script de extração das opiniões de usuários do site TripAdvisor Brasil[5]

```
# Instala as bibliotecas necessárias
# install.packages(c("rvest"))

# Carrega as bibliotecas necessárias
library(rvest)

## Loading required package: xml2

# Cada página subsequente difere, no meio do endereço (orXX-), pelos termos
do seguinte vetor "looping". Dependendo do caso, deve-se alterar o número d
e termos
looping<-c("","or10-","or20-","or30-","or40-","or50-","or60-","or70-","or80
-","or90-","or100-","or110-","or120-")
n<-length(looping)
dataset <- data.frame()

for(i in looping){

  # Endereço da página do restaurante cujos dados deverão ser extraídos
  url <- paste ("https://www.tripadvisor.com.br/Restaurant_Review-xxxxxxx-x
xxxxxxxx-Reviews-",i,"Nome-Do-Restaurante.html#REVIEWS",sep="")

  # Variável com o endereço da página
  reviews <- url %>%
    read_html() %>%
    html_nodes("#REVIEWS .innerBubble")

  # ID do review (gerado pela página)
  id <- reviews %>%
    html_node(".quote a") %>%
    html_attr("id")

  # Título do review criado pelo cliente
  titulo <- reviews %>%
    html_node(".quote span") %>%
    html_text()

  # Pontuação da experiência definida pelo cliente em escala de 10 a 50
  pontuacao <-  html_nodes(x = reviews, css = ".rating.reviewItemInline") %
>%
    html_children() %>%
    html_attr("class")
  pontuacao <- pontuacao[grepl("ui",pontuacao)]
  pontuacao <- as.numeric(substr(pontuacao,nchar("ui_bubble_rating bubble_1
0")-1,nchar("ui_bubble_rating bubble_10")))

  # Data do review
```

[5] Os nomes das empresas foram removidos

```
  datas <- reviews %>%
    html_node(".rating .ratingDate") %>%
    html_attr("title")

  # Comenbtário em texto criado pelo cliente
  texto <- reviews %>%
    html_node(".entry") %>%
    html_text()

  # Limpa os \n dos comnetário que, neste caso, representam o 'enter' a cad
a final de linha e podem bagunçar o dataframe
  textoreview <- gsub("\n", "", texto)

  # Cria um dataframe com tudo
  temp.dataset <- data.frame(titulo, pontuacao, datas, texto)
  dataset <- rbind(dataset,temp.dataset)

}

# Salvando o dataframe com os dados extraídos
write.csv(dataset, "restaurante.csv")
```

APÊNDICE B – Lista final de *stop words* aplicada na pesquisa

de
a
o
que
e
do
da
em
um
para
é
com
não
uma
os
no
se
na
por
mais
as
dos
como
mas
foi
ao
ele
das
tem
à
seu
sua
ou
ser
quando
muito
há
nos
já
está
eu
também
só
pelo
pela
até
isso
ela
entre
era
depois
sem
mesmo
aos
ter
seus
quem
nas
me
esse
eles
estão
você
tinha
foram
essa
num
nem
suas
meu
às
minha
têm
numa
pelos
elas
havia
seja
qual
será
nós
tenho
lhe
deles
essas
esses
pelas
este
fosse
dele
tu
te
vocês
vos
lhes
meus
minhas
teu
tua
teus
tuas
nosso
nossa
nossos
nossas
dela
delas
esta
estes
estas
aquele
aquela
aqueles
aquelas
isto
aquilo
estou
está
estamos
estão
estive
esteve
estivemos
estiveram
estava
estávamos
estavam
estivera
estivéramos
esteja
estejamos
estejam
estivesse
estivéssem
estivessem
estiver
estivermos
estiverem
hei
há
havemos
hão
houve
houvemos
houveram
houvera
houvéramos
haja
hajamos
hajam
houvesse
houvéssemos
houvessem
houver
houvermos
houverem
houverei
houverá
houveremos
houverão
houveria
houveríamos
houveriam
sou
somos
são
era
éramos
eram
fui
foi
fomos
foram
fora
fôramos
seja
sejamos
sejam
fosse
fôssemos
fossem
for
formos
forem
serei
será
seremos
serão
seria
seríamos
seriam
tenho
tem
temos
tém
tinha
tínhamos
tinham
tive
teve
tivemos
tiveram
tivera
tivéramos
tenha
tenhamos
tenham
tivesse
tivéssemos
tivessem
tiver
tivermos
tiverem
terei
terá
teremos
terão
teria
teríamos
teriam
q
vc