FUNDAÇÃO MINEIRA DE EDUCAÇÃO E CULTURA
UNIVERSIDADE FUMEC
MESTRADO EM SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO

JÚLIO CÉSAR BARBOSA

Mineração de texto: uso de técnicas de processamento de linguagem natural para suporte à geração de projeções baseadas em opiniões do consumidor

Belo Horizonte - MG
2018

JÚLIO CÉSAR BARBOSA

Mineração de texto: uso de técnicas de processamento de linguagem natural para suporte à geração de projeções baseadas em opiniões do consumidor

Dissertação apresentada ao Programa de Mestrado em Sistemas de Informação e Gestão do Conhecimento da Fundação Mineira de Educação e Cultura – FUMEC, como requisito parcial para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento.

Área de concentração: Sistemas de Informação e Gestão do Conhecimento.

Linha de Pesquisa: Tecnologia e Sistemas de Informação.

Trilha de pesquisa: T4 – Recuperação da Informação.

Orientador: Prof. Dr. Luiz Cláudio Gomes Maia.

Belo Horizonte - MG

2018

Dados Internacionais de Catalogação na Publicação (CIP)

B238m Barbosa, Júlio César, 1980 -

Mineração de texto: uso de técnicas de processamento de linguagem natural para suporte à geração de projeções baseadas em opiniões do consumidor / Júlio César Barbosa. – Belo Horizonte, 2018.

67 f. : il. ; 29,7 cm

Orientador: Luiz Cláudio Gomes Maia

Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento), Universidade FUMEC, Faculdade de Ciências Empresariais, Belo Horizonte, 2018.

1. Processamento de linguagem natural (Computação). 2. Processo decisório - Brasil. 3. Inteligência competitiva (Administração) - Brasil. 4. Tecnologia - Brasil. I. Título. II. Maia, Luiz Cláudio Gomes. III. Universidade FUMEC, Faculdade de Ciências Empresariais.

CDU: 65.01:001

Ficha catalográfica elaborada pela Biblioteca Universitária-FUMEC

UNIVERSIDADE
FUMEC

Dissertação intitulada "**Mineração de Texto: Uso de técnicas de Processamento de Linguagem Natural para suporte à geração de projeções baseadas em opiniões do consumidor**" de autoria de Júlio César Barbosa, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Luiz Cláudio Gomes Maia – Universidade FUMEC
(Orientador)

Prof. Dr. Fabrício Ziviani – Universidade FUMEC
(Examinador Interno)

Prof. Dr. Fernando Hadad Zaidan – IETEC
(Examinador Externo)

Priscila Silveira de Lacerda, Me. – SMART Inteligência Empresarial
(Consultor *Ad Hoc*)

Prof. Dr. Fernando Silva Parreiras
Coordenador do Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC

Belo Horizonte, 11 de setembro de 2018.

☐ **REITORIA**
Av. Afonso Pena, 3880 - Cruzeiro
30130-009 - Belo Horizonte, MG
Tel. 0800 0300 200
www.fumec.br

☐ **CAMPUS**
Rua Cobre, 200 - Cruzeiro
30310-190 - Belo Horizonte, MG
Tel. (31) 3228-3000
www.fumec.br

Agradecimentos

Fazer um mestrado é uma experiência engrandecedora, e que somente em seu tempo certo pode ser aproveitada em sua plenitude.

É um tempo de dedicação, introspecção e que exige muita escolha, onde muitas vezes impactam até os que estão a nossa volta. Tendo isso, não é somente justa, mas merecida as palavras de agradecimento aos que nos cercam e que fazem juntos esta caminhada.

Ao meu estimado orientador Professor Luiz Claudio Gomes Maia, pela serenidade, assertividade e principalmente pela sensibilidade em entender limitações, quer pessoais ou profissionais, que por muitas vezes impactaram minha pesquisa. Agradeço pelo seu tempo dedicado e por cada direcionamento que recebi.

A minha querida esposa Patrícia, pelo companheirismo, paciência e pelas inúmeras dicas e contribuições sobre esse encantador mundo acadêmico e as particularidades da pesquisa *stricto sensu.*

Ao meu filhinho Lucas, que chegou para iluminar minha vida bem no meio desta pesquisa.

A minha mãe, fonte deste querer aprender que sempre me acompanhou.

A minha família, pela força e motivação.

Muito obrigado!

Resumo

Não há precedentes sobre a frequência e volume com que as pessoas publicam informações virtualmente, cujo maior atrativo é que grande parte dos usuários da internet posta suas opiniões, expectativas, elogios e frustrações, concebendo, assim, vasta base de dados primários sobre diversos produtos e serviços. O propósito deste trabalho foi o uso de técnicas de processamento de linguagem natural, por meio do *Natural Language Tool Kit* (NLTK), para extração de características ou aspectos, pré-processamento do *corpus* e classificação da polaridade, criando então uma base de dados indexados, que facilitou a análise dos dados, apoiando assim a geração de projeções de oportunidades. A base de dados foram as opiniões postadas em *sites* de venda de celulares. Foi possível, por meio da análise dos dados tratados e indexados, projetar oportunidades de melhoria do produto, como, por exemplo, melhorar o aspecto bateria, pois em produtos concorrentes esse aspecto é recorrentemente elogiado ao longo dos anos. Também foram projetadas oportunidades de melhoria do processo de venda, identificado em aspectos orbitantes, como foi o caso de opiniões negativas sobre determinado revendedor. A expectativa é de que essas projeções possam dar suporte às empresas em seu processo decisório, mais precisamente porque poderão servir de insumo ao processo de inteligência competitiva. Buscou-se, dessa forma, gerar um resultado que trouxesse diferencial competitivo, usando fontes de dados abertas, em formato primário e coletadas de maneira ética e legal.

Palavras-chave: Mineração de opinião. Extração de características ou aspectos. Análise de sentimentos. Processamento de Linguagem Natural.

Abstract

This work aims to show up Natural Language Processing techniques in order to support the construction of indexed database which will facilitate analysis and support the generation of projections of business enhancements, based on opinions posted over websites, regarding cell phones reviews. Was possible through the analysis of data results, infer opportunities for product and process improvements. Opportunities have also been designed to improve the sales process, identified in orbiting aspects, as seen in case of negative opinions about a specific vendor. Those projections are expected to be able to support companies in their decision-making process, more precisely because it can serve as input to competitive intelligence process. The purpose was to generate a competitive differential, through data adding value, using as a corpus, open data in a primary format and collected ethically and legally.

Keywords: Opinion mining. Characteristics extraction. Sentiment analysis. Natural Language Processing.

Lista de Figuras

Lista de Quadros

Lista de Tabelas

Lista de Abreviaturas e Siglas

AS	Análise de sentimentos
BPMN	*Business Process Model and Notation*
CRF	*Conditional random fields*
HTML	*Hyper Text Markup Language*
IC	Inteligência Competitiva
LIWC	*Linguistic Inquiry and Word Count*
MD	Mineração de dados
MO	Mineração de opinião
MT	Mineração de Texto
NLTK	*Natural Language Tool Kit*
PLN	Processamento de linguagem natural
POS	*Part-of-speech*
SASA	*Sail Ail Sentiment Analyser*
SCIP	*Society of Competitive Intelligence Professionals*
SVM	*Support Vector Machine*
TI	Tecnologia da informação
TXT	Formato de texto simples
URL	*Uniform Resource Locator*

Sumário[1]

[1] Este trabalho foi revisado de acordo com as novas regras ortográficas aprovadas pelo Acordo Ortográfico assinado entre os países que integram a Comunidade de Países de Língua Portuguesa (CPLP), em vigor no Brasil desde 2009. E foi formatado de acordo com as Instruções para Formatação de Trabalhos Acadêmicos – Norma APA, 2017.

1 Introdução

Nos últimos anos tem-se observado significativo aumento em relação à frequência com que as pessoas interagem por meio de mídias sociais e pelos mais variados meios informacionais digitais, não só consumindo informações, mas também expondo suas opiniões. Diante disso, uma base de dados sem precedentes aumenta a cada dia. É possível facilmente ler opiniões, comparações, resenhas e comentários sobre os mais diversos assuntos, produtos e serviços, em grupos de discussão, *sites* especializados e *blogs.* Entretanto, esta vasta base de opiniões e comentários se apresenta de maneira não estruturada, onde já em 2001, Chen (2001), indicou em seu trabalho que 80% do conteúdo da *web* era informação textual.

No trabalho de Berry e Kogan (2010), é citado que com o avançar da tecnologia, o volume de soluções dedicadas ao armazenamento de informações não estruturadas também evoluiu. Neste ecossistema, onde “é estimado que aproximadamente 85% dos dados em rede é está em formato não estruturado” Berry et al. (2010, p. 183), onde é possível que muitos dados que apresentem padrões de comportamento, tendências e preferências estejam sendo também armazenados. Porém, estas características possam estar ocultas, pois em seu formato natural de armazenamento os dados sem nenhum tratamento não possuem valor agregado, do ponto de vista de suporte ao processo de decisão, pois não farão sentido ao consumidor da informação. Muitas vezes, este dado em forma natural nada mais é que uma opinião de uma pessoa expressada textualmente e descrita de forma coloquial.

Liu (2012), apresenta uma interpretação de que uma opinião tem origem em uma ação de alguém, que a apresenta de forma negativa ou positiva em seu discurso, e que associa esta opinião a uma determinada característica de outra pessoa ou objeto.

Inúmeros trabalhos relacionados ao tratamento de opiniões, com intuito de adição de valor a esses dados disponibilizados abertamente na internet, tem sido realizados nos mais variados segmentos de mercado. Como visto em Gao, Tang, Wang e Yin (2018), onde trataram opiniões postadas na internet sobre restaurantes. Buscaram agregar valor aos resultados gerados de forma que determinado restaurante pudesse entender melhor a percepção de seu empreendimento em relação aos concorrentes, na visão do cliente. Já nos trabalhos de Akhtar, Zubair,

Kumar e Ahmad (2017) e Moertini, Kevin e Satyadi (2017), a mesma abordagem foi feita, mas com opiniões referentes ao segmento de hotelaria.

É possível encontrar, inclusive, trabalhos relacionados à análise de opiniões postadas, que buscam identificar padrões que revelem tendências suicidas, como é o caso do trabalho de Birjali, Beni-Hssane, e Erritali (2017), onde coletaram dados do *Twitter* e os trataram com abordagens de aprendizado de máquina e análise semântica, no intuito de busca por indicativos deste tipo de tendência.

Segundo Liu (2012), a forma com que as pessoas podem se expressar textualmente é ilimitada, ou seja, dados textuais podem ser gerados, agrupados e disseminados de múltiplas formas. Uma intervenção manual, no sentido de coleta, tratamento

A análise dessas informações não estruturadas em formato de opiniões, com o intuito de busca por conhecimento ou quaisquer tendências que possam apoiar uma empresa na avaliação do que seu público diz sobre seu produto, é tarefa não produtiva e insignificante em termos de amostra se considerado o volume encontrado na *internet*. Além disso, opiniões podem ser encontradas em diversos formatos ou granularidade, isto é, podem ser classificadas em nível de documento, sentença ou entidades e aspectos:

a) Documento: infere que há uma polaridade geral que se aplica a todo o documento referente àquela entidade em questão;
b) sentença: granularidade média, em que cada sentença do texto pode expressar uma opinião sobre determinada característica daquela entidade única;
c) entidade e aspecto: nível de granularidade alta, cujo documento pode conter mais de uma entidade e várias opiniões sobre cada aspecto desta, definindo, assim, pares entidade x aspecto.

Neste trabalho será considerada a análise em nível de entidade e aspecto, uma vez que serão analisadas a polaridade da opinião sobre aquele aspecto mapeado, pois poderá ser detectado mais de um em cada sentença.

Como exemplo, no texto *“*aquele celular Samsung custa caro, mas a bateria é muito boa”, tem-se a entidade “Celular Samsung” e dois aspectos: preço e bateria. É um indicado em um contexto de polaridade negativa e outro positiva.

Xu, Liao, Li & Song (2011) afirmam que a volumetria de dados encontrados para montagem do *corpus* torna-se um ativo de extremo valor quando a empresa aplica o processo de inteligência competitiva (IC). Inicialmente porque se trata de uma abordagem estratégica que sustenta o processo de tomada de decisão das empresas e também embasa a gestão de riscos, que tem como um de seus fatores restritivos a falta ou o insuficiente volume de informações.

É nesse ponto que este experimento pretende contribuir, especificamente no uso de ferramentas e técnicas de processamento de linguagem natural (PLN), para tratar computacionalmente uma base de dados em larga escala, que será o *corpus* deste trabalho. *Corpus*, neste contexto, configura-se como um conjunto de documentos que versam sobre o mesmo tema e que compuseram a presente amostra de dados. Nesse caso, serão as opiniões postadas em *sites* de venda de celulares *online*, para que possa ser possível identificar quais são os aspectos mais citados de acordo com o público, classificar sua polaridade e, dessa forma, dar suporte à tomada de decisão e embasar a gestão de riscos. Isso porque são dados primários em larga escala de uma base pública e foram adquiridos de forma legal e ética, como preconiza o processo de IC, buscando, assim, gerar diferencial competitivo para a empresa.

1.1 Justificativa

As citações na literatura, no que se refere à definição de IC, em mais ou em menos detalhamento, na maior parte das vezes convergem em um ponto comum, de que se trata de uma abordagem processual de coleta de dados abertos. Ou seja, dados públicos e seu devido tratamento devem buscar por algum conhecimento que traga à empresa vantagem competitiva e suporte à tomada de decisão e que este processo seja ético e legal (Canongia, 1998; Coelho, Dou, Quonian & Silva, 2006; *Society of Competitive Intelligence Professionals* - SCIP, 2018).

Sites de comércio eletrônico fornecem abundante fonte de dados sobre produtos por eles comercializados. Esses dados, ou, nesse caso, as opiniões dos clientes, são expressados das mais diversas formas: comparando um produto com outro, comparando uma única característica daquele produto em relação ao do concorrente e elogiando ou reclamando de determinado evento durante a compra (ex.: atraso na entrega, embalagem danificada). Essas opiniões poderiam ser tratadas de maneira estruturada, de forma a se extrair tendências importantes que contribuíssem

com a estratégia das empresas (Sun, Luo & Chen, 2017) para projeção de novos produtos, ajuste no portfólio, ajuste em apenas uma das características ou aspectos daquele produto ou mesmo melhorando o processo desde a venda até a entrega do produto, tornando-se, assim, cada vez mais aderentes às necessidades do cliente. Dado esse cenário, é necessário, antes de tudo, criar uma base única, ou *corpus*, gerado por meio das mais diversas fontes, conforme Tang, Qin & Liu (2015).

Como a percepção e decisão do consumidor é influenciada pelas suas emoções (Bollen, Mao & Zeng, 2011), sua exteriorização, nesse caso, por meio de informação textual, é a base primária de dados sobre determinado assunto. Chega-se então à questão central deste estudo, que trata do uso de técnicas de PLN para captar, armazenar, tratar e distribuir informações que possam gerar uma base de informações que, após analisada pelo profissional de IC da empresa, possa gerar projeções baseadas nos comentários postados ou *reviews* sobre o produto e seus aspectos (fonte de dados primária). Essas projeções poderão ser usadas pela empresa como insumo ao processo de IC, pois é a abordagem que busca gerar conhecimento e, por sua vez, irá dar suporte ao processo de tomada de decisão.

O produto celular foi escolhido neste trabalho, por ser tema de publicação já feita nessa mesma linha de pesquisa, como pode ser encontrado em Xu *et al.* (2011), e também por se tratar de bem comum, acessível para compra por diversos meios, além de possuir considerável montante de empresas concorrentes. Dessa forma, fornece oportunidade para que o consumidor possa ter percepção do que uma empresa oferece de melhor do que outra, tanto em termos de aspectos do produto, como aspectos da transação comercial, como preço, tempo de entrega, garantias, pós-vendas, entre outros. Estes últimos são classificados como aspectos orbitantes do produto.

1.2 Lacuna a ser explorada

Maia e Souza (2010) consideram que nos últimos anos um volume de informações abundante tem sido registrado em bases de dados nas diversas áreas do conhecimento e sob inúmeras formas (numéricas, textuais, imagens, etc.), como ilustrado pela Figura 1. E Baeza-Yates & Ribeiro-Neto (1999) levam em conta que o impacto gerado pela popularização da internet e seu uso como intermediador de compras *online* estão nas características que as informações são persistidas nesse

ambiente. Tanto aqueles autores como esses entendem que a *web* tornou-se um espaço de negócios onde as pessoas consomem e compartilham tanto bens como informações, nesse caso no formato de postagens como opiniões sobre esses bens.

Figura 1. Postagens em mídias sociais durante 1 minuto.
Fonte: https://www.digitalinformationworld.com/2018/05/infographic-internet-minute-2018.html#.

Se for possível, por meio desses dados não estruturados, distribuídos nas páginas em formato de opiniões e disponíveis nesse ambiente de compartilhamento não pré-formatado, projetar oportunidades de melhoria no produto ou processo, tendo como base a percepção demonstrada pelo público em relação a um produto ou ao do concorrente, essas projeções poderiam servir de insumo ao processo de IC que, por sua vez, dará suporte à tomada de decisão das empresas.

Serão usadas as *reviews* dos consumidores postadas no *site* da empresa *Amazon.com*. São fonte de dados abertos, ou seja, disponíveis também ao concorrente, primária (é a opinião do próprio consumidor, descrita por ele diretamente) e em grande volume, tornando-se assim a matéria-prima do experimento.

1.3 Problema de pesquisa

Quais projeções, no sentido de antecipação de situação futura ou melhorias no produto ou processo de venda, poderiam ser feitas após o devido tratamento por meio da aplicação de técnicas de PLN, em um *corpus* de opiniões sobre celulares vendidos *online* e que possam dar suporte à tomada de decisão e agregar valor à estratégia (IC) para empresas que operam com vendas de celulares pela internet?

1.4 Contribuição da pesquisa

A aplicação de técnicas de PLN e pesquisa descritiva, que segundo Gil (2008, p. 42) “têm como objetivo primordial a descrição das características de determinada população ou fenômeno ou, então, o estabelecimento de relações entre variáveis”, contribui com o processo IC, dando suporte à tomada de decisão. E como fonte primária de dados as opiniões postadas na internet pelos consumidores que compraram celulares naquela loja *online*.

Essas opiniões são comumente postadas em redes sociais (*facebook, twitter*, etc.), *sites* de venda dos produtos (*amazon.com, americanas.com*, etc.) e até mesmo em *sites* centralizadores de serviços (trivago.com.br, decolar.com, etc.). Entende-se que esta pesquisa nas bases de dados de opiniões dos clientes pode ser uma valiosa fonte de insumo para as empresas, sustentando, inclusive, o processo de gestão de risco.

A adequação ao programa de pós-graduação da FUMEC em Sistemas de Informação e Gestão do Conhecimento é focada em pesquisa aplicada e com foco na multidisciplinaridade.

Esta pesquisa, além de tratar do uso de ferramentas e técnicas de PLN, busca criar, por meio dessas ferramentas, conhecimento que possa subsidiar a estratégia de empresas. É de natureza multidisciplinar e está inserida no campo de pesquisa em Sistemas de Informação, quando usou ferramentas de Tecnologia da Informação (TI) e buscou aperfeiçoamento da lógica do código que tratou da etapa de *crawling.* Com isso, intenta melhorar o mecanismo de recuperação da informação para montagem do *corpus* e da Administração de Empresas, quando propõem, por meio de mecanismos de TI, especificamente técnicas de PLN, a adição de valor aos dados

que poderão sustentar o processo de tomada de decisão empresarial, portanto, aderente à proposta do programa.

1.5 Objetivos geral e específicos

O objetivo geral foi recuperar, processar e disponibilizar, por meio de técnicas de PLN, dados com valor agregado, ou seja, com informações qualitativas e quantitativas que, quando analisadas e interpretadas por uma pessoa com o intuito de busca por padrões ou tendências, possam ajudar na geração de projeções de melhorias. Essas projeções podem indicar quais os aspectos ou características do produto ou processo que podem ser melhorados ou alterados e servem de insumo ao processo de tomada de decisão. Buscou-se, assim, adicionar valor à informação coletada por meio ético e legal, de uma base primária e pública, conforme preconizado pelo processo de IC.

Como objetivos específicos os seguintes:

a) Automatizar a extração das opiniões, por meio de técnicas de *crawling*, visando otimizar o tempo despendido na etapa de construção do *corpus*.
b) Identificar características ou aspectos e definir polaridade por meio das técnicas de PLN, usando a biblioteca *Natural Language Tool Kit* (NLTK) desenvolvida na linguagem *Python*, visando detectar elementos que possam ser melhorados no produto ou processo.
c) Disponibilizar dados indexados com informações qualitativas e quantitativas, como frequência de citação e polaridade do termo, com o propósito de apoiar o profissional de IC para que os analise, na busca por algum padrão no sentido de gerar projeções de melhoria.

1.6 Estrutura do documento

Visando organizar este trabalho, seu conteúdo foi dividido nesta introdução, na qual foram apresentados o contexto, justificativa, lacuna a ser explorada, problema de pesquisa, contribuição da pesquisa e objetivos, geral e específico; na fundamentação teórica, com as informações conceituais sobre os temas abordados, trabalhos relacionados, algumas técnicas e o resultado da coleta de informações para

montagem do *corpus* deste trabalho; na metodologia, que descreve a abordagem processual implementada, o detalhamento das etapas de desenvolvimento; e nos resultados encontrados, com o resumo das contribuições alcançadas, trabalhos futuros e limitações da pesquisa.

2 Fundamentação Teórica

Neste capítulo são apresentados alguns trabalhos de autores clássicos relativos ao tema desta dissertação, além de trabalhos recentemente publicados. Serão também mais bem contextualizados os demais temas abordados, como inteligência competitiva, processamento de linguagem natural e mineração de textos, sendo mais bem explicados e exemplificados com trabalhos clássicos e recentemente também publicados.

2.1 Trabalhos relacionados

O trabalho de Xu *et al.* (2011) registra uma pesquisa referente à análise de sentimentos, precisamente a análise comparativa de opiniões, tendo como base opiniões sobre aspectos ou características de celulares e que foram comparados aos resultados da abordagem tradicional.

A análise comparativa de opiniões versa sobre a natureza de uma sentença quando expressa uma ordem de relação entre dois ou mais conjuntos de entidades, quando está explícita a citação de uma ou mais características em comum desses elementos. Como exemplo, na seguinte sentença “o celular Samsung tem uma bateria melhor que o da Apple e da Motorola” está clara a seguinte relação de comparação: [melhor (bateria Samsung), (Apple), (motorola)].

O trabalho de Xu utilizou essa técnica como forma de sustentar o processo de inteligência competitiva da empresa proprietária daquele celular, oferecendo uma abordagem mais precisa que pudesse fornecer projeções mais assertivas sobre as opiniões de seus produtos quando em comparação aos concorrentes. O resultado dessa nova abordagem indicou ter precisão 3,41% superior à abordagem tradicional, usando *conditional random fields* (CRF), que é um método de modelagem estatística aplicada em reconhecimento de padrões e aprendizado de máquina, mas que não captura a relação do sentimento em nível de entidade, pois desconsidera a polaridade em nível de palavra. Assim, propuseram o uso do CRF em dois níveis, em que se considerasse a direção da polaridade de cada uma das entidades mapeadas na sentença.

Outro trabalho, proposto por Kauer (2016), refere-se a dois pontos principais relacionados à mineração de opiniões, à análise de sentimentos baseados em

aspectos e à atribuição de polaridade. Ele desenvolveu um método que identifica aspectos e entidades nos textos, usando técnicas de PLN e aprendizado de máquina. É um método para atribuição de polaridade baseado num *ranking* previamente determinado por aprendizado de máquina, que pode ser aplicado em um *corpus* com ruído, que mesmo assim alcançou resultados próximos das abordagens tradicionais, mais complexas em termos de implementação.

Além das opiniões sobre produtos ou aspectos do produto, há crescente aumento das avaliações sobre serviços, como os relativos às redes hoteleiras e restaurantes, em quesitos como tempo de entrega, preço, atendimento, etc. Esse assunto é tratado no trabalho de Siqueira (2010), que desenvolveu um *software* de análise de sentimentos automatizado e implementou alguns processos, como pré-processamento, mapeamento de indicadores, identificação dos substantivos frequentes, mais relevantes e remoção dos não relacionados. O resultado foi a concentração, em um único produto, de uma solução que percorre alguns processos para análise de sentimentos baseada em opiniões sobre serviços.

No trabalho de Valsan, Sreepriya & Nitha (2017) foi proposta uma abordagem que foi classificada como híbrida, por ter envolvido abordagens supervisionadas e não supervisionadas. Eles trabalharam com um *corpus* relacionado a opiniões postadas sobre câmeras digitais, coletadas em *sites* de venda.

Para o método supervisionado, o trabalho de Valsan *et al*. (2017) utilizou o *Support Vector Machine* (SVM) para realizar a tarefa de análise dos sentimentos. Essa abordagem baseada em aprendizado de máquina realiza a tarefa de classificar e separar as diferentes categorias, treinando esse conjunto de dados e inferindo os sentimentos que são classificados em positivo, negativo e neutro com base no conjunto de dados treinados previamente.

Na etapa não supervisionada foi utilizado o algoritmo K-Means, que se trata de um dos mais frequentemente usados na mineração de dados e geração de estatísticas. Depois de pré-processar o conjunto de opiniões, foram recuperadas as palavras e sua frequência de ocorrências naquela amostra. Esses dados são usados como entrada para o algoritmo K-Means que, por sua vez, forma diferentes *clusters* baseados nos conjuntos de palavras, como pode ser visto na Figura 2.

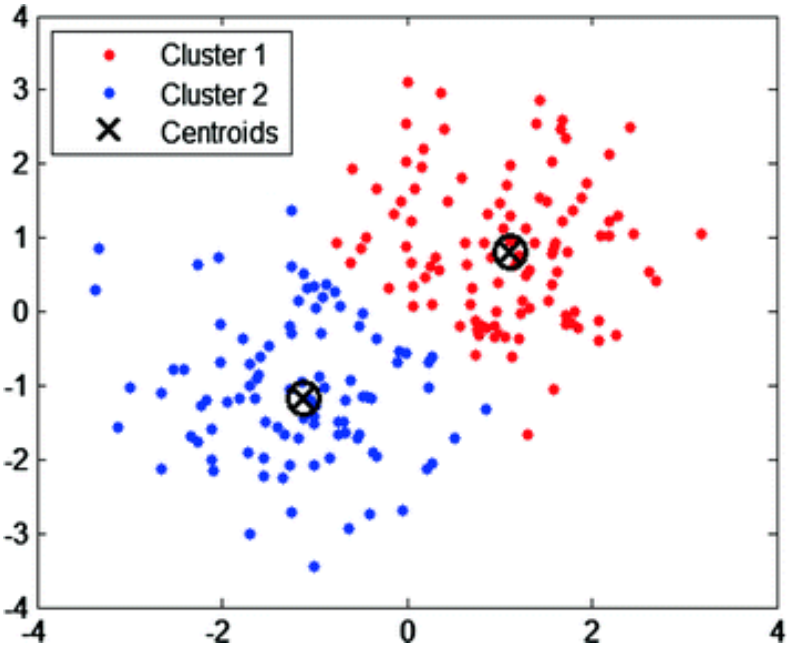

Figura 2. Resultado do *cluster* feito pelo K-Means.
Fonte: Valsan A., Sreepriya C.T., & Nitha L. (2017). Social media sentiment polarity analysis: a novel approach to promote business performance and consumer decision-making. *In*: S. Dash, K. Vijayakumar, Panigrahi B., & S. Das (eds) Artificial intelligence and evolutionary computations in engineering systems. *Advances in Intelligent Systems and Computing, 517,* Springer, Singapore.

Esses dados tratados e posteriormente estruturados estatisticamente foram a base para permitir entender melhor as tendências dos consumidores, gerando, dessa forma, diferencial competitivo para empresa. O resultado da abordagem híbrida foi que o uso combinado trouxe resultados clusterizados dos termos mais citados nas opiniões, usando o K-Means, e que depois tiveram previamente classificada sua polaridade pelo SVM, fornecendo, com isso, ao consumidor dados com valor qualitativo para decisão de qual câmera comprar (Valsan *et al*., 2017).

O trabalho de He, Wu, Yan, Akula & Shen (2015) sugere uma solução para análise competitiva de mídia social, por meio de análise de sentimentos, e que foi usada para tratar inteligência de *marketing* específica de alguns setores industriais. A intenção foi identificar quais empresas de mídia estão liderando naquele tópico referente a determinada indústria.

A Figura 3 exemplifica a estrutura proposta para análise competitiva de mídia social, segmentada pela indústria a que se refere:

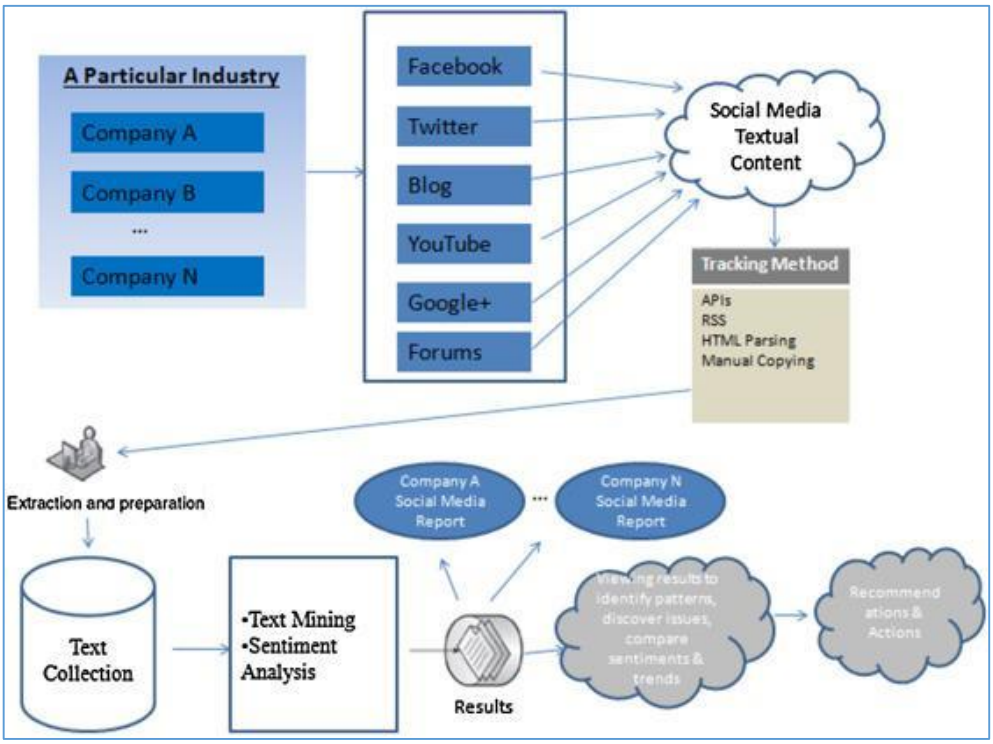

Figura 3. *Framework* proposto para análise competitiva de mídia social.
Fonte: He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(issue 7), 801-812.

Primeiramente, foram escolhidas algumas empresas que se destacam em seu campo de atuação, como Costco, Walmart, Kmart, Kohl's, e The Home Depot, na indústria de varejo. Em seguida, por meio do *Twitter*, He *et al*. (2015) pesquisaram informações relacionadas a essas empresas e que serviram de massa de dados para uma ferramenta de TI desenvolvida por eles, para análise competitiva de mídia social orientada a negócios, denominada VOZIQ. Primeiramente, o VOC3, que é módulo de relatórios sobre *benchmark* de sentimentos da ferramenta VOZIQ, que reúne menções aos clientes, concorrentes e clientes dos concorrentes, classifica essas menções em mídia social em várias categorias por meio de métodos de classificação supervisionada e não supervisionada, para identificar categorias de assuntos. Por fim, classifica o sentimento dentro dessas categorias.

O VOZIQ também comparou a participação de diferentes plataformas de mídia social nas menções sobre as empresas. Essa funcionalidade permite a comparação do desempenho da marca da empresa em vários *sites* de mídia social.

Uma das atividades mais importantes quando se trabalha com a análise de sentimentos relativa a opiniões sobre produtos é a devida classificação da polaridade da opinião sobre o aspecto de acordo com o contexto em análise.

Kansal & Toshniwal (2014) propõem uma abordagem de análise de sentimentos baseada em aspectos, por meio de processamento de linguagem natural e técnicas de mineração de opinião, que considerou o contexto em que o aspecto foi citado para classificação da polaridade. Ou seja, considerou, inclusive, a polaridade das demais sentenças com mesmo aspecto no *corpus* em análise.

Inicialmente foi usado um dicionário de termos disponível *online* que serviu para classificação dos termos, independentemente do contexto em que se encontravam. Em seguida, foram aplicadas regras linguísticas para definir polaridade aos termos (palavras de opinião) de acordo com o contexto. Em seguida, os autores apresentaram as opiniões já devidamente classificadas quanto à polaridade, porém, com o seu par aspecto-palavra de opinião. Puderam, assim, demonstrar que a polaridade pode mudar de acordo com o âmbito em análise. Os resultados demonstraram que a abordagem de classificação considerando o cenário do termo teve precisão superior quando apenas analisada de forma geral.

2.2 Inteligência competitiva

A IC, estudada e definida sob o enfoque de processo, tem em suas citações: "objetiva agregar valor à informação, fortalecendo seu caráter estratégico, catalisando, assim, o processo de crescimento organizacional. Nesse sentido, a coleta, tratamento, análise e contextualização de informação permitem a geração de produtos de inteligência" (Canongia, 1998, p. 2-3).

Já Coelho *et al*. (2006), descrevem a IC como sendo um processo sistemático de captar, classificar, analisar, controlar e distribuir informações relacionadas ao mercado interno e externo, portanto, fluido e definido em etapas claras e bem delimitadas, podendo, portanto, ser aplicado de forma sistemática na empresa.

A *Society of Competitive Intelligence Professionals* (SCIP) define o conceito de IC também na linha de um processo sistematizado, claro e bem definido, porém, adiciona a esse entendimento os aspectos legais e éticos da atividade e por meio dessas práticas éticas e legais o processo deve ser inserido no fluxo de tomada de decisões estratégicas da empresa, agregando, então, valor ao negócio.

Tendo-se agora não só o aspecto processual da IC, Valentim, Lenzi, Cervantes, de Carvalho, Garcia, Catarino e Tomaél (2003) declaram ser esse um procedimento que investiga o âmbito onde a empresa se encontra, buscando atenuar riscos, além,

é claro, de monitorar também internamente seu ambiente, “visando o estabelecimento de estratégias de curto, médio e longo prazo” (Valentim *et al*., 2003, p. 1). Chegaram à conclusão de que “a prospecção e o monitoramento informacional desenvolvem e apoiam a inovação tecnológica, atividade essencial para a competitividade organizacional” (Valentim *et al*., 2003, p. 16).

Sendo considerado um ferramental estratégico, torna-se, portanto, “crucial a disponibilização de informações relevantes sobre questões estratégicas para os negócios, de maneira a subsidiar os decisores das organizações quanto à avaliação das mudanças de mercado e a identificação de tendências, de novos entrantes de substitutos e de oportunidades [...]” (Lopes, De Muÿlder & Judice, 2012, p. 215).

Ainda sob a ótica da IC como catalisadora da estratégia corporativa, observa-se o fato de que a geração e o consumo massivo de informações são constantes, como apresenta Vickery (2018, parágrafo 16) “o cidadão, então, está num ato de consumo ou até um momento de lazer – mas, na verdade, está produzindo, fornecendo informações que têm grande potencial econômico para quem é capaz de pegar, analisar, vender”. Logo, a empresa que “pega esses dados tem nas mãos o combustível do mundo atual” (Vickery, 2018, parágrafo 16).

Quando inserido o aspecto TI na abordagem da IC, alguns autores posicionam-se quanto à sua essencialidade, afirmando que “a empresa que melhor perceber as aplicações das tecnologias emergentes às suas operações, e que puder usar eficazmente a informática aos processos decisórios, terá maior vantagem competitiva” (Teixeira Filho, 2000 como citado em Valentim *et al*., 2003, p. 12).

Nos trabalhos de Amarouche, Benbrahim & Kassou (2015) são apresentadas duas abordagens processuais sobre as etapas que a IC precisa percorrer. A primeira, conhecida como PCMAC, refere:

a) *Plan and Prioritize:* o trabalho é planejado, os recursos e os tópicos-chave são definidos, etc.;
b) *capture:* fase de coleta da informação;
c) *manage:* a informação é filtrada, separada e compilada;
d) *analyse* a análise é feita, efetivamente;
e) *communicate:* fase final, cujos achados são compartilhados com os grupos de interesse da organização.

A segunda abordagem considera:

a) Identificação das necessidades de IC: os tópicos-chave para aplicação da IC são definidos;
b) aquisição de informação competitiva: coleta de dados de diversas fontes de interesse;
c) organização, armazenamento e recuperação: fase de organização e armazenamento da informação;
d) análise da informação: fase principal da IC, na qual a informação é transformada em conhecimento;
e) disseminação: distribuição do conhecimento descoberto.

Ambas as abordagens podem ser claramente distribuídas no tempo, sendo cada fase claramente separada da outra e a relação predecessora e sucessora deve ser mantida.

Em relação ao tempo, é possível dizer, sobre o processo de IC, “que o principal objetivo é preparar as ações futuras com base nas saídas da fase de análise da informação. Assim sendo, o eixo do tempo, durante este processo de tomar decisão no tempo certo, se mostra muito importante”, segundo Amarouche *et al*. (2015). Esse entendimento pode ser mais bem entendido na Figura 4:

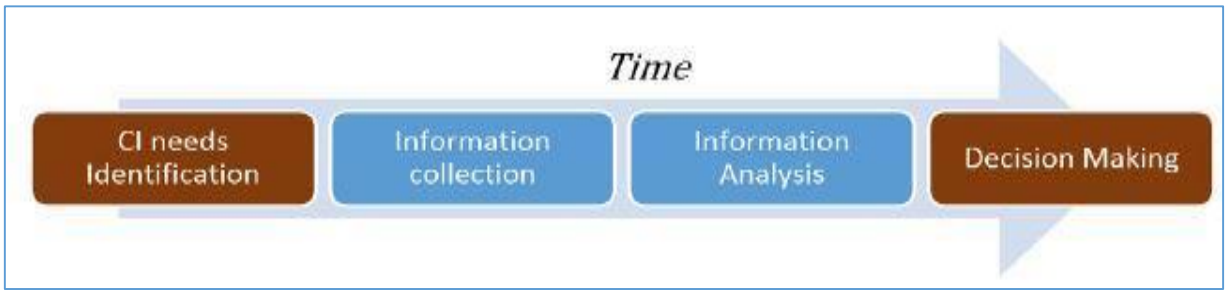

Figura 4. Processo de IC em relação ao tempo.
Fonte: Amarouche, K., Benbrahim, H., & Kassou, I. (2015). Product opinion mining for competitive intelligence. *Procedia Computer Science, 73*, 358-365.

A partir dessas definições, tem-se a oportunidade, neste trabalho de uso de técnicas de PLN, que são ferramentas de TI, para gerar projeções que, por sua vez, irão gerar “inteligência acionável, de forma a favorecer a conquista de seus objetivos táticos ou estratégicos”, conforme descrito por Garcia (2018, p. 131).

O processo de IC será, assim, beneficiado, pois como ferramenta de reconhecido aspecto estratégico à organização e com características processuais, portanto, possível de ser modularizado em termos de etapas ou componentes, pode ser associado a métodos e técnicas com base computacional, fazendo uso da TI como catalisador, usando, nesse caso, abordagem de mineração de texto (MT), especificamente aplicando técnicas de PLN (Garcia, 2018).

Essa similaridade das duas abordagens, IC e MT, podem ser mais bem entendidas se observada uma comparação visual das etapas macros que compõem cada uma, como mostra a Figura 5:

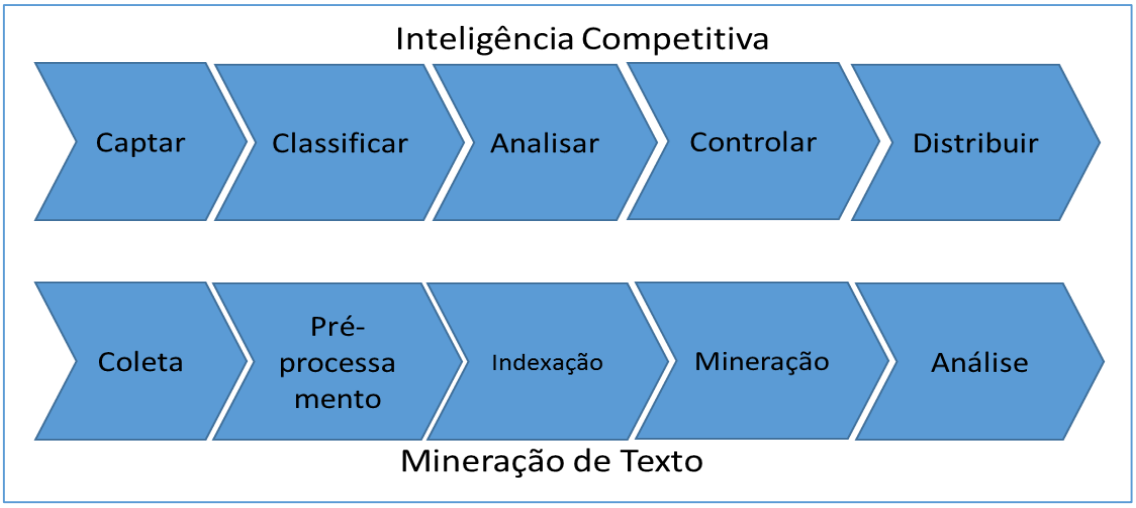

Figura 5. Etapas dos processos de IC e MT.
Fonte: adaptado de Coelho, G. M., Dou, H,. Quonian, L., Silva, C. H. (2006). Ensino e pesquisa no campo da inteligência competitiva no Brasil e a cooperação franco-brasileira. *Revista Hispana de La Inteligencia Competitiva - Puzzle*. España, año 6, (23), 12-19; Garcia, A. E. G. (2018). Inteligência competitiva: considerações sobre a prática no ambiente empresarial brasileiro. *Revista Inteligência Competitiva, 8*(1), 127-168; Aranha, C. N., & Vellasco, M. (2007). *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, RJ.; Amarouche, K., Benbrahim, H., & Kassou, I. (2015). Product opinion mining for competitive intelligence. *Procedia Computer Science, 73*, 358-365..

2.3 Processamento de linguagem natural

O PLN é uma área interdisciplinar na qual se incluem a Informática, Matemática, Ciência da Informação, Psicologia e Linguística. Estudam-se a representação e interpretação automática do dado não estruturado, ou melhor explicando, da informação falada ou escrita por humanos. Essa é uma área em que os pesquisadores são orientados na busca por entendimento de como humanos geram, fazem uso e compartilham informação, direcionando seus esforços para a

construção de *softwares* que possam manipular a linguagem natural o mais próximo da interpretação humana (Andrade, 2018).

Essa área de pesquisa é normalmente aplicada, mas não se limita à recuperação da informação, ao reconhecimento de voz, ao processamento de texto e à tradução automática (Chowdhury, 2003).

Vieira e Lima (2001) conceituam o PLN como uma abordagem orientada ao desenvolvimento de programas de *software* com capacidade de tratamento, interpretação e geração de informações em linguagem humana.

Na visão de Turban, Leidner, Mclean e Wetherbe (2010), o PLN refere-se ao processamento da linguagem humana e sua forma de interação entre homem e máquina, independentemente do idioma em que está seja descrito.

Como uma das mais complexas etapas do PLN está o tópico conhecido como resolução de correferência, que versa sobre como interpretar quando pronomes ou substantivos descobertos naquela amostra textual em análise se referem à mesma entidade em estudo. A abordagem mais tradicional para tratamento desse item é por meio do uso de aprendizado supervisionado, em que recebe um par de parâmetros e infere se são correferentes. Em seguida, quando novas amostras são inseridas, essa função é novamente aplicada, porém sobre todas as entidades presentes, formando assim um novo agrupamento. Dessa forma, segue-se o aprendizado, formando conjuntos cada vez maiores, mas com uma pré-classificação já inferida. É possível observar, no trabalho de Stoyanov & Cardie (2006), técnicas que utilizam essa abordagem ainda na etapa de extração das características.

Outro importante tema relacionado ao PLN, diz respeito ao agrupamento de sinônimos. Como dito, normalmente os substantivos e sua frequência de citação tendem a indicar qual entidade à qual se refere aquela amostra de texto. Entretanto, é muito frequente que termos sinônimos possam ser usados no referenciamento da entidade em questão. No caso de celulares, o termo “bateria” ou “carga” poderiam se referir ao mesmo aspecto e muitas vezes dentro da mesma sentença. A busca pelo agrupamento dos sinônimos dentro da amostra traz melhoria do processo de classificação na etapa de pré-processamento da amostra. O uso de dicionários como o *WordNet,* que, inclusive, é usado pelo módulo do NLTK, responsável pela extração de características, contribui consideravelmente para o agrupamento dos sinônimos. No entanto, não resolve completamente a questão, uma vez que, dependendo do domínio em análise, um termo pode ter significados distintos.

No tratamento de dados não estruturados, ou textos, alguns dos procedimentos de PLN são dedicados à preparação do *corpus* em estudo, para que computacionalmente possam ser interpretados. Alguns dos procedimentos incluem a tokenização e posterior análise sintática do texto.

Como pode ser visto na Figura 6, esse processamento é requerido para que, computacionalmente, seja possível classificar sintaticamente uma sentença:

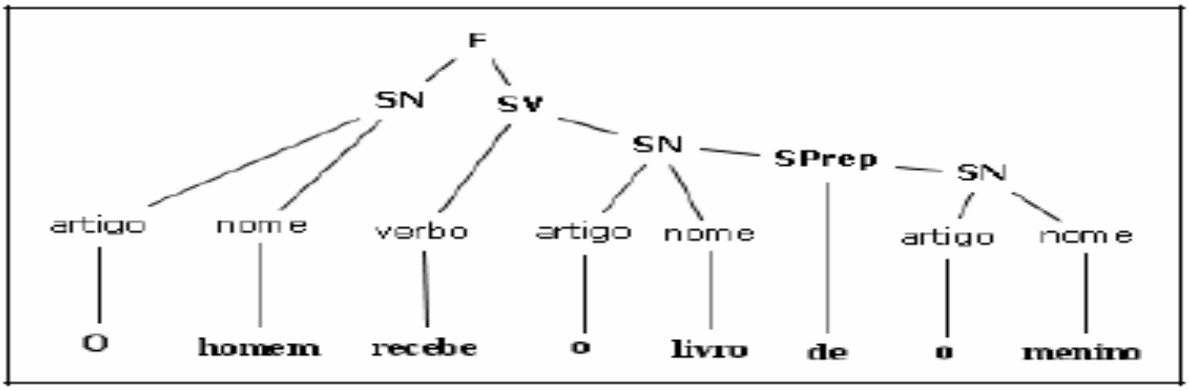

Figura 6. Análise da estrutura da sentença.
Fonte: Souza, R. R. (2005). Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.

Essa marcação das palavras da sentença, ou comumente conhecida com *POS tagging,* fornece identidade a cada termo em relação a seu valor sintático. Substantivos normalmente indicam as entidades às quais a frase faz referência. Na Tabela 1 relacionam-se conjuntos de *POS tagging*.

Tabela 1

Conjunto de PS *Tagging*

Nome	Descrição	Nome	Descrição
CC	Conjunções coordenativas	PRP$	Pronome possessivo
CD	Numeral cardinal	RB	Advérbio
DT	Delimitador	RBR	Advérbio comparativo
EX	*"there"* existencial	RBS	Advérbio superlativo
FW	Palavra estrangeira	RP	Palavras inflexivas
IN	Conjunções subordinativas	SYM	Símbolo
JJ	Adjetivo	TO	*"to"* como preposição
JJR	Adjetivo comparativo	UH	Interjeição
JJS	Adjetivo superlativo	VB	Verbo, forma básica
LS	Marcador de item	VBD	Verbo, passado
MD	Verbo auxiliar	VBG	Verbo, gerúndio
NN	Substantivo singular	VBN	Verbo, particípio
NNP	Substantivo próprio	VBP	Verbo, presente
NNPS	Substantivo próprio plural	VBZ	Verbo, 3ª pessoa singular
NNS	Substantivo plural	WDT	"wh"determinante
PDT	Pré-determinante	WP	"wh"pronome
POS	Indicador possessivo	WP$	"wh"pronome possessivo
PRP	Pronome pessoal	WRB	"wh"adverbial

Nota. Fonte: Kauer, A. U. (2016). Análise de sentimentos baseada em aspectos e atribuições de polaridade. Dissertação (Mestrado em Administração) - Porto Alegre: Universidade Federal do Rio Grande do Sul., p. 24.

Algumas técnicas de PLN são comumente utilizadas com a abordagem conhecida como baseada em léxico, ou seja, baseada não só em um vocabulário, mas também em construtos, como algumas palavras isoladas e também em sentenças completas, que por sua vez contribuem para detecção das relações semânticas do *corpus* em tratamento. Essa técnica é reconhecida como uma das mais simples, uma vez que determinados modelos utilizam, inclusive, o caractere de espaço como sendo um elemento, consequentemente, gerando uma marcação a esse elemento na fase de POS. Entretanto, em idiomas como o mandarim, o espaço não necessariamente é sempre caracterizado apenas como separador de palavras. Nesse caso, outra abordagem deve ser usada.

2.4 Mineração de textos

Minerar textos, diferentemente do processo de mineração de dados (MD), concentra-se na extração de padrões de informações tendo como insumo grandes volumes de dados textuais não estruturados ou semiestruturados (Maia & Souza, 2010). Já o segundo trabalha na descoberta de padrões em dados de uma massa pré-normalizada, ou seja, estruturada, segundo Han, Pei e Kamber (2011). Ambas as

abordagens utilizam procedimentos de pré-processamento, entretanto, a MD mantém foco na normalização e associações dos dados, enquanto a MT busca a identificação dos elementos que possam caracterizar indubitavelmente o alvo sob análise (Hu & Liu, 2004). Posteriormente, ambas voltam o foco para a descoberta de padrões que possam gerar conhecimento e sustentar o objetivo a que se propõem.

Minerar textos pode ser entendido como um conjunto de procedimentos com foco em se obter informação de relevância para o propósito a que se destina, tendo como dados a serem analisados os textos escritos em linguagem natural, ou linguagem humana (Hotho, Nürnberger & Paaß, 2005).

Segundo Aranha e Passos (2006), minerar textos não é o mesmo que apenas usar um mecanismo de busca. Este segundo pressupõe que o usuário já tenha consciência de qual informação quer encontrar. Portanto, minerar textos é buscar o conhecimento ainda desconhecido e que possa trazer diferencial competitivo, quer seja ao indivíduo ou a uma empresa.

Esse processo é naturalmente complexo e pode ser subdividido nas seguintes etapas vistas na Figura 7, conforme estruturado por Aranha e Vellasco (2007):

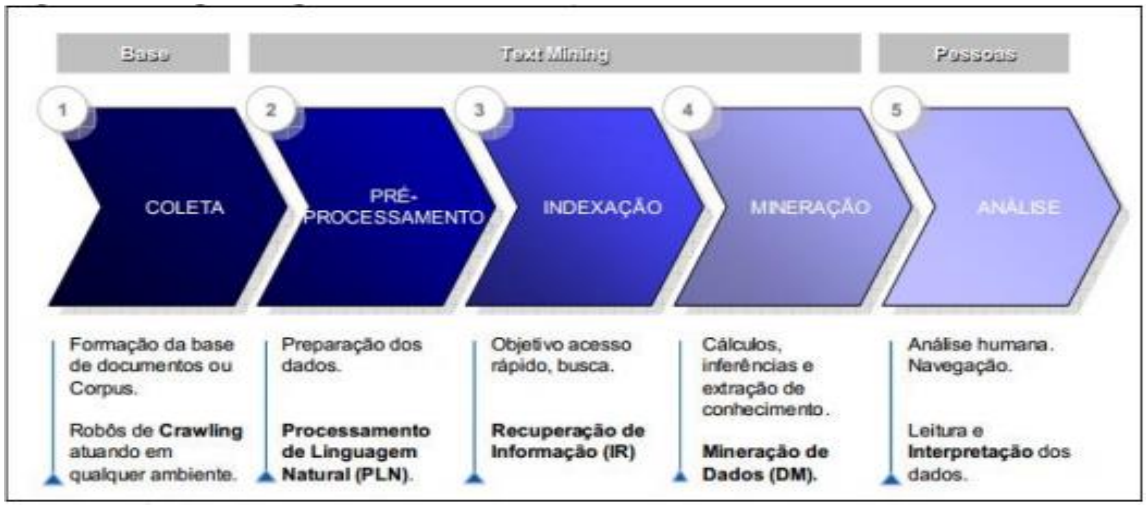

Figura 7. Fases do processo de mineração de textos.
Fonte: Aranha, C. N., & Vellasco, M. (2007). *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, RJ.

Na fase de **coleta** executa-se a aquisição de dados, estando aí a origem de tudo. É etapa de importância para o desenvolvimento, pois a qualidade da massa amostral, sua relevância para a pesquisa a que se destina, bem como a legalidade

com que é conseguida tem relação direta com o resultado que se busca. Dessa etapa gera-se o *corpus* inicial a ser processado.

A fase de **pré-processamento** tem a finalidade de preparar a informação original de forma a torná-la adaptável ao método de mineração. Aqui, a natureza de cada processo ou técnica de MT influencia diretamente cada etapa, deixando-a ou mais ou menos complexa do ponto de vista de transformação dos dados textuais.

Na fase de **mineração** são aplicados algoritmos e procedimentos estatísticos visando extrair informações relevantes da base de dados que já possam endossar a interpretação humana, sendo esta última etapa conhecida como análise dos dados. E é onde são realizadas as interpretações e visualizações dos resultados extraídos.

Por fim, na fase de **análise** encontra-se a base de informações já trabalhada, indexada e pronta para interpretação humana, ou seja, é o alvo de construção e produto-fim para interpretação e busca de conhecimento para geração das projeções. O que se busca nessa fase é identificar algum padrão ou indicativo de tendências nos dados apresentados que possam ser usados a favor da empresa, seja salientando a necessidade de melhorias ou mesmo reconhecendo algum aspecto positivo já existente, que deve ser mantido, uma vez que possui reconhecido valor pelo cliente.

2.4.1 Análise de sentimentos ou mineração de opinião

São conceitos similares e referem-se à abordagem de identificar emoções e opiniões em bases de informação não estruturadas - textos.

A análise de sentimentos (AS) ou mineração de opinião (MO) são subáreas do processo de MT e versam sobre a forma de extração subjetiva de informações dentro de fontes de dados textual (Liu, 2012). É possível classificar essas extrações de acordo com o sentimento que o autor da informação pretendeu passar, definindo assim a sua polaridade (positivo, negativo ou neutro).

Não é abordagem recente, sendo citada em pesquisas com pouco mais de 10 anos, como visto em Wiebe, Wilson, Bruce, Bell & Martin (2004) e Liu (2012) entre outros. Na visão Pang & Lee (2008), em 2001 houve o marco inicial das pesquisas aplicadas neste tema, tornando-se, assim, área de grande disseminação.

Tsytsarau & Palpanas (2012) afirmam que a MO refere-se à questão de identificação de uma opinião expressa relacionada a determinado tópico e avaliação da polaridade dessa opinião, se positiva ou negativa, por exemplo. A MO fornece

detalhada visão das emoções contidas e explicitadas no *corpus*, permitindo o processamento posterior de dados, buscando, inclusive, identificação de opiniões contraditórias. Dessa forma, fica bem evidente que a qualidade dos resultados dessa fase do processamento é fator crucial para o êxito de todas as tarefas posteriores, tornando-se uma etapa de extrema importância.

Esse processo tem seu início na etapa de coleta de informação, que pode ser por meio da extração de dados de *blogs*, mídias sociais, *sites* de notícias, *sites* de *reviews* sobre produtos, etc. Essa informação, ou muitas vezes uma frase de opinião de algum consumidor sobre um produto, precisa ser pré-processada para que seja possível sua classificação de forma mais automatizada, ou seja, identificar as características ou aspecto em questão e posteriormente a classificação de sua polaridade. Cada uma dessas etapas citadas pode ser realizada por técnicas ou abordagens distintas, como visto na Figura 8:

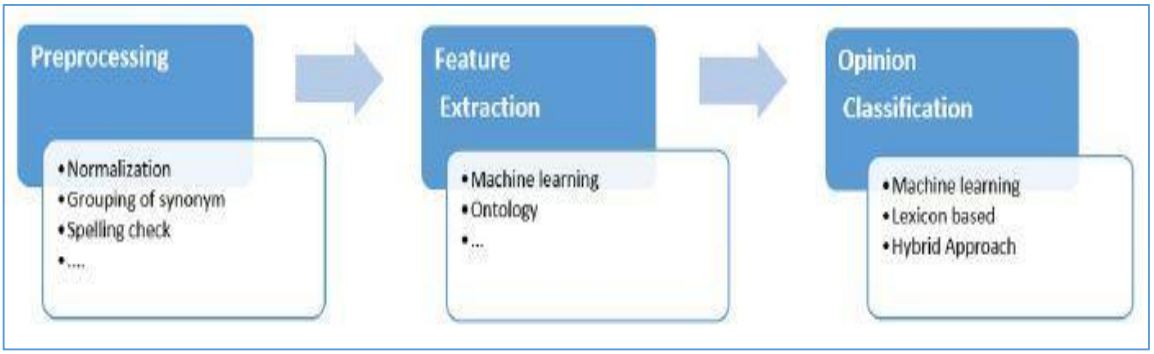

Figura 8. Etapas e abordagens da MO.
Fonte: Aranha, C. N., & Vellasco, M. (2007). *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, RJ.

Alvo da opinião: Kauer (2016, p. 18) enfatiza que “o alvo é definido por sua posição única de início e final em um texto” e “este alvo pode não estar explícito no texto”, “usando pronomes ou correferências textuais”. Algumas abordagens para extração do alvo da opinião são citadas por Hu & Liu (2004), Qiu, Liu, Bu & Chen (2011) e Pontiki, Galanis, Papageorgiou, Manandhar & Androutsopoulos (2015), como:

a) Extração baseada em substantivos frequentes, em que por meio de análise gramatical são identificados os substantivos mais frequentes e lhes são atribuídos pesos;

b) por meio da relação alvo x sentimento, em que também por meio de analisador gramatical as relações de dependência são buscadas para evidenciar o alvo e as palavras que denotam sentimentos;
c) por meio de aprendizado supervisionado, em que modelos são usados para determinar se aquela opinião se refere àquele aspecto.

Polaridades dos sentimentos: o *Natural Language Tool Kit* (NLTK) é um conjunto de ferramentas para processamento de linguagem natural que foi inicialmente desenvolvido pela Universidade da Pensilvânia em conjunto com o curso de Linguística Computacional daquela instituição em 2001 (Bird & Loper, 2004). É modular e seus componentes podem ser usados isoladamente de forma a se adaptar ao requisito de processamento que se pretende daquele *corpus*.

Ele disponibiliza alguns módulos para o tratamento de cada fase do PLN. Cada um desses módulos trata de uma etapa, desde a remoção das conhecidas, como *stopwords* (preposições, pronomes, artigos, etc.), pois são componentes textuais que não possuem valor informacional para a análise do contexto da sentença.

No trabalho de Symeonidis, Effrosynidis & Arampatzis (2018), o NLTK foi usado como uma das técnicas na etapa de remoção de *stopwords,* na fase de pré-processamento dos dados. Essa pesquisa tratou da comparação de 16 técnicas de pré-processamento em dois conjuntos de dados extraídos do *Twitter*, visando fazer a análise de sentimentos. As demais técnicas usadas constam no Quadro 1:

Quadro 1

Técnicas de pré-processamento

Number	Técnica de pré processamento
0	Básico (Remover cadeias de caracteres Unicode e ruído)
1	Outro (substituir URLs e menções de usuários)
2	Substitua Gírias e Abreviações
3	Substituir Contrações
4	Remover números
5	Substituir Repetições de Pontuação
6	Substituir Negações por Antônimos
7	Remover pontuação
8	Manipulando Palavras Capitalizadas
9	Minúsculas
10	Remover palavras irrelevantes
11	Substituir palavras alongadas
12	Correção ortográfica
13	POS-Tagging
14	Lemmatizing
15	Stemming
16	Negações

Fonte: adaptado de Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*.

No caso específico de dados extraídos do *Twitter*, o uso da técnica de remoção de *stopwords* não foi muito satisfatório. Uma das justificativas é que textos dessa plataforma são feitos na maioria das vezes por jovens, que tendem a usar textos curtos, com muitas gírias e muitas palavras para referenciarem a si mesmos, que são classificadas pela técnica como *stopwords*, e com isso removidas.

Há um módulo que trata da definição da orientação do texto, ou seja, de sua polaridade, indicando ser esta positiva, negativa ou neutra. Esse tratamento comumente faz uso de um dicionário léxico, com palavras previamente mapeadas como indicativas de polaridade positiva ou negativa.

Um comum exemplo é a associação negativa da sentença, quando palavras de negação são encontradas, como neste exemplo no caso do idioma inglês: *“don't", "never", "nothing", "nowhere", "none", "not", "hasn't","hadn't","can't"*, etc.

2.4.2 Extração de aspectos ou categorias

O uso de técnicas de extração de características ou aspectos visa à busca pelas entidades e seus aspectos em uma base de dados não estruturados. Isso permite recuperar e gerar informações, descobrir padrões e projetar tendências. A partir desse ponto, surgiram diversas pesquisas voltadas para a seleção de características ou aspectos, para gerar conhecimento (Pak & Paroubek, 2010) e a identificação de diferentes tópicos (Hu & Liu, 2004).

Classicamente, a atividade de extração de características pode ser classificada em dois grandes grupos. O primeiro é constituído d os que fazem uso de estatística e linguística e para tal usam classes gramaticais ou análise da frequência de palavras, no sentido de busca por elementos ou termos mais relevantes naquela amostra, e com isso conseguem classificar com mais acurácia a classe gramatical do termo, analisando para isso sua distribuição sintática e morfológica. Essa etapa é comumente conhecida na literatura como *POS Tagging*, que tem como resultado a etiquetagem do termo, para posterior análise. Essas etiquetas marcam cada termo com sua classe, como, por exemplo, numeral, verbo, substantivo, preposição, etc.

Ainda dentro dessa primeira abordagem, há a análise de frequência de palavras, que infere que as características mais significativas normalmente são substantivos, pois comumente são repetidas dentro do cenário em análise. Portanto, os termos que mais aparecem têm mais probabilidade de serem classificados como sendo a característica em questão.

Existe também o segundo grupo, que é baseado em aprendizado de máquina, que faz uso de algoritmos que buscam identificação das polaridades, por meio da classificação dos termos e busca por padrões.

O motivo pelo qual um consumidor elogia uma característica ou aspecto de um produto (exemplo: tamanho da tela do celular) pode não necessariamente indicar que ele aprecia tudo relacionado àquele produto. Considerando esse comportamento, “a extração de características é uma das tarefas mais difíceis e menos propensas à automação da análise de sentimentos”, e por isso “os trabalhos que tratam a extração de características não seguem nenhuma abordagem consensual” (Siqueira, 2010, p. 29). Percebe-se, então, representativo volume de soluções que implementam essa tarefa, porém com abordagens muitas vezes ligeiramente diferentes.

Um exemplo de trabalho relacionado que usou essa técnica foi apresentado por Kumar & Abirami (2018), que desenvolveram um *framework* para classificar opiniões baseadas em aspectos ou categorias, baseado no método de correlação de Spearman, que avalia de forma monotônica a relação entre variáveis contínuas. O termo monotônico refere-se à característica de duas ou mais variáveis que têm de mudar de forma conjunta, mas não necessariamente a uma taxa constante no tempo.

O *framework* desenvolvido no trabalho de Kumar & Abirami (2018) consistiu nos seguintes elementos: o processo de coleta de dados, pré-processamento, extração de aspectos, análise de opinião baseada em aspectos, classificação de opinião baseada em aspectos e classificação de sentimentos e avaliação. A Figura 9 exemplifica essa proposta:

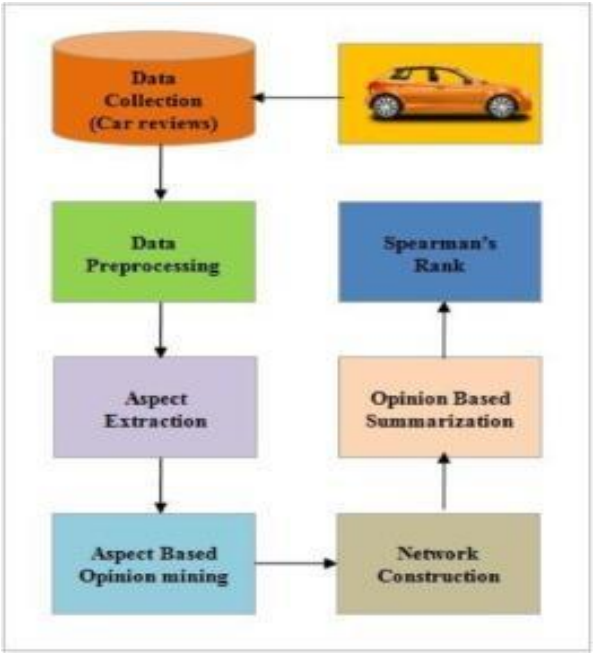

Figura 9. *Framework* proposto para classificação baseada no método de Spearman.
Fonte: Kumar, A., & Abirami, S. (2018). Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method. *Information Sciences*, *460-461,* 23-41..

Como entidade a ser analisada, os autores escolheram carros produzidos entre 2007 e 2009 e o *corpus* a ser analisado foi composto pela coleta de *reviews* feitas em *sites* especializados, nos quais foram coletadas 42.230 *reviews* relativas a aproximadamente 150 modelos de carros. Após as etapas de pré-processamento, as *reviews* foram classificados por polaridade.

Como produto, uma relação de 11 veículos foi selecionada e as polaridades médias de cada aspecto em comum entre os veículos foi medida, gerando o resultado da Figura 10:

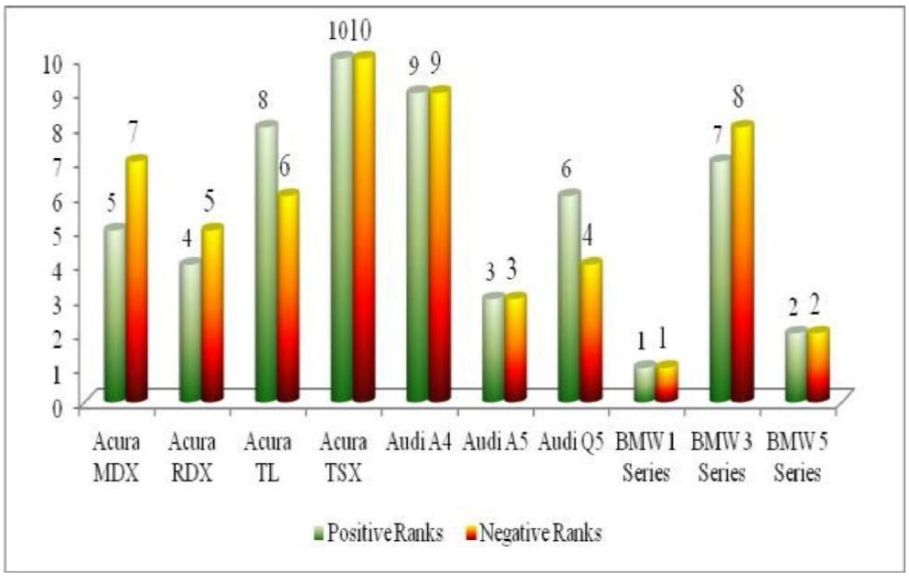

Figura 10. Polaridade média dos aspectos em comum.
Fonte: Fonte: Kumar, A., & Abirami, S. (2018). Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method. *Information Sciences*, *460-461,* 23-41.

O resultado da pesquisa desses autores revelou que houve tendência à classificação mais positiva entre as feitas individualmente para cada entidade, que foram feitas sendo ponderadas pelo método de Spearman.

3 Procedimento Metodológico

Serão detalhados os procedimentos e etapas percorridos para responder ao problema de pesquisa, bem como atender aos objetivos específicos desta dissertação.

3.1 Classificação da pesquisa

Este trabalho, pela sua natureza, classifica-se como aplicada, pois se pretende adquirir novos aprendizados e gerar conhecimento (Mascarenhas, 2010). Na visão de Barros e Lehfeld (2000), sendo aplicada, motiva-se pela busca da produção de conhecimento, pois dessa forma poderá contribuir de maneira prática com demandas reais à nossa volta. Teve objetivos descritivos, portanto, é natural que ambas as técnicas - qualitativa e quantitativa - fossem aplicadas, pois são inerentes a essa abordagem (Mascarenhas, 2010).

3.2 Resumo das etapas

Como foram trabalhados dados não estruturados nesta investigação, a abordagem de MT foi aplicada, implementada por meio da aplicação das técnicas de PLN. Foi extraído, armazenado e tratado o *corpus* de forma a descobrir padrões e com isso gerar projeções que possam subsidiar a construção de conhecimento. Este trabalho baseou-se na proposta de Aranha e Vellasco (2007) e Garcia (2018), cujas etapas do processo serão detalhadas no item 3.3. Na Figura 11 é apresentado o fluxo de atividades a serem realizadas em cada fase.

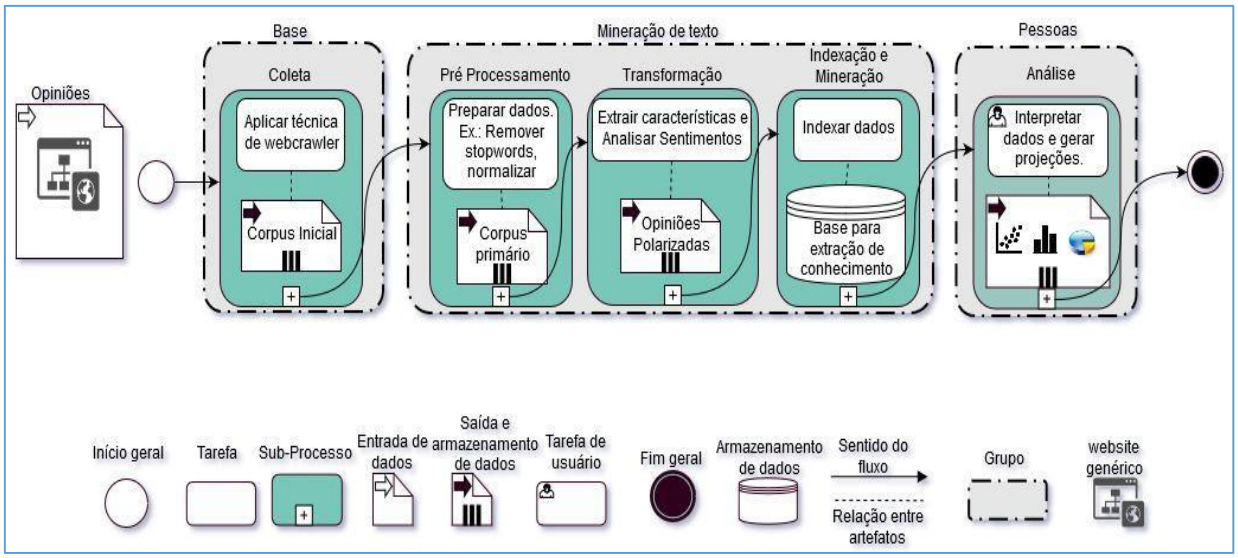

Figura 11. Etapas do processo de MT: fluxo BPMN.
BPMN: *Business Process Model and Notation.*
Fonte: adaptado de Aranha, C. N., & Vellasco, M. (2007). *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, RJ; Garcia, A. E. G. (2018). Inteligência competitiva: considerações sobre a prática no ambiente empresarial brasileiro. *Revista Inteligência Competitiva, 8*(1), 127-168..

3.3 Desenvolvimento

3.3.1 Base

Nessa etapa (subprocesso: coleta e montagem do *corpus*), foi feita a coleta das informações que foram analisadas, nesse caso, opiniões dos consumidores postadas na internet. Nela foi aplicada a técnica de *webcrawler,* que consistiu em automatização do processo de extração de informações de páginas *web*, conforme marcação (*tag*), pré-delimitada. Especificamente, foi utilizada a biblioteca *jSoup* (*jsoup.org,* 2018*)*, da linguagem Java, cujo propósito e funcionamento, assim como o detalhamento da técnica de *webcrawler,* serão apresentados a seguir. Foram adicionadas algumas funcionalidades que serão explicadas.

Na Figura 12 é reportado o código-fonte desenvolvido:

```
import java.io.IOException;
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import org.jsoup.select.Elements;
import java.io.FileWriter;
import java.io.PrintWriter;

public class Raspagem{
public static void main(String[] args) throws IOException{
     FileWriter arq = new FileWriter("C:\\Users\\Julio\\corpus.txt");
     PrintWriter gravarArq = new PrintWriter(arq);
        try {
            for (int j=2; j < 60; j++){
               String url = "https://www.amazon.com/Samsung-Galaxy-S8-Unlocked-64GB/product-reviews/B06Y\r"+
                            "14T5YW/ref=cm_cr_arp_d_paging_btm_"+j+"2?ie=UTF8&reviewerType=all_reviews&pageNumber="+j+"";
               Document doc = Jsoup.connect(url).get();
               Elements elements = doc.select("span[data-hook=\"reviewBody\" class=\"a-size-base review-text\"]");

               for(int i = 0; i < elements.size(); i++) {
                  gravarArq.printf(elements.get(i).text()+" | ");
               }
            }//fim do FOR externo
            arq.close();
         }
         catch (IOException e) {
            e.printStackTrace();
         }
     }
}//Fim da Classe Raspagem
```

Figura 12. Código-fonte usado para extração de opiniões e montagem do *corpus*.
Fonte: o próprio autor.

Foi passada a *Uniform Resource Locator* (URL), que nada mais é que o endereço do *site* onde foi feita a coleta das opiniões, exatamente na subpágina onde as opiniões são postadas. O método *jSoup.connect()* recebe essa URL como parâmetro e se conecta ao *site*. Em seguida, identificou-se no código-fonte do *site* que a *tag* de marcação HTML referente às opiniões era <*span*> e que o atributo dessa *tag* que armazena as opiniões propriamente ditas era o <*data-hook="reviewBody" class="a-size-base review-text*">**.**

Esse atributo foi passado ao método *select(),* que teve como resultado vários elementos, nesse caso, o texto referente às opiniões daquela página específica. Em seguida, usando um comando de repetição, foram salvas em um arquivo .txt todas as opiniões contidas naquelas páginas.

Como cada página do *site* traz apenas algumas opiniões, foi automatizado o mecanismo de busca por meio da inclusão de uma variável na URL, no exato trecho do endereço que indica a página, em seguida passada ao método *jSoup.connect(),* que é incrementado a cada *loop*. Dessa forma, percorreram-se todas as páginas de postagens com opiniões.

A seguir, apresenta-se uma amostra do resultado obtido da extração de algumas opiniões, usando-se a técnica de *web crawling* que compôs o *corpus*:

Nice phone. It does what a smart phone should. Nice size screen for the few games my wife uses while waiting in doctor's offices. The only problem I had was trying to add it to my AT&T account. I already had another phone with AT&T.I ordered a SIM card for the phone and attempted to activate it. No go. I called to get help and they told me the phone was not 4G and I would have to buy a new one from them. They kept saying it wasn't 4G and it couldn't be added to their system.They sell a Samsung Galaxy J3.I was so unhappy with AT&T I went with Consumer Cellular with the Galaxy J3 and moved my LG Flex to them also,The phone shows LTE when connected to Consumer Cellular, as I understand it, that tells me it is 4G. I Need a Phone that I Can Dump All of My Pictures and Music From The Internal Memory to The Memory Card.! That's what I Thought that I Had when I Purchased This One. But AT&T Tells Me that I Have to Use The Cloud. What's the Use in Having a Memory Card For? Good Quality, Functions like a charm. Got this phone since I dropped my Samsung 3 in water and ruined it. Wasn't sure on purchasing it, but I am not sorry. Good phone, easy to use and lightweight. Recommend this phone for people who likes Samsung.

Conforme Siqueira (2010), o termo *Web Crawler* significa um programa que navega em páginas *web*, de forma estruturada, e cria cópias dessas páginas para posteriormente serem analisadas de forma metódica e estruturada. Essa técnica também indexa essas páginas para prover buscas mais eficientes.

Considerando que o volume de opiniões postadas aumenta diariamente e nem sempre as informações são encontradas formatadas de maneira adequada à nossa necessidade, essa técnica de *crawler* reduz a árdua tarefa de coletar essas informações manualmente.

A biblioteca *jSoup* da linguagem Java implementa essa técnica de *crawler* e foi utilizada neste trabalho. As fontes de informação que compuseram o *corpus* tiveram origem em *sites* de revenda de celulares *online,* como, por exemplo, a *www.amazon.com.*

Não foi utilizado o próprio *site* da empresa fabricante do produto, para evitar que as opiniões disponibilizadas pudessem ter sido selecionadas de forma parcial, retirando assim precioso insumo que é o dado primário e direto do consumidor final.

Essa biblioteca *jSoup* da linguagem Java foi usada no projeto para manipulação de conteúdo em formato *HyperText Markup Language* (HTML), mas conhecido como *crawling*. Com ela foi possível ler e extrair informações de arquivos HTML. Ela realiza

o *parser*, que faz a análise das páginas *web* definidas, de forma a se extrair as informações necessárias àquela necessidade. Para definir qual informação deveria ser extraída, foi apontado no código dessa biblioteca a *tag* de marcação que identifica a informação que precisa ser extraída (*jsoup.org*, 2016), nesse caso, as opiniões postadas.

O *jSoup* se conecta à página definida, por meio do método *Jsoup.connect()*. Ele recebe como parâmetro a URL do *site*. Esse método executa duas funções, a primeira é se conectar ao *site* e fazer o *download* deste para que seja analisado; e realizar o *parsing*, retornando um objeto do tipo *Document.* Esse objeto *Document* armazena a página de forma segmentada, delimitando suas marcações HTML, para que se possa executar a extração necessária. Quando é feita a chamada ao método *title()* sobre o objeto *Document*, há como retorno o título do *site* que está armazenado no objeto *Document*.

Se for necessário acessar elementos (*tags*) específicos, deve-se usar o método *select()*, do objeto *Document*. Esse método retorna um objeto do tipo *Elements,* que se trata de um conjunto de objetos *Element* que representou cada informação retornada daquela *tag* especificada.

Outra possibilidade, que, inclusive, foi a forma aplicada neste trabalho, foi a extração de informações específicas dentro de uma *tag*, ou seja, do atributo dessa *tag*. Foi usado novamente o método *select(),* mas desta vez, além de passar somente a *tag* como parâmetro, foi passado o atributo desse *tag* que foi extraído.

3.3.2 Mineração de texto

Nessa etapa (subprocesso: pré-processamento), o tratamento dos dados desse *corpus* inicial foi por meio da aplicação de técnicas de PLN, usando as soluções disponibilizadas pelo NLTK. A primeira etapa, o pré-processamento, teve a finalidade de reduzir a sentença extraída na etapa de coleta a uma unidade mínima que fizesse sentido semântico.

Essa atividade é conhecida como remoção dos *stopwords.* Basicamente, esse procedimento foi feito para se identificar o que podia ser eliminado do texto, ou seja, tudo que não tinha função informacional. Normalmente são classificados como *stopwords* as conjunções, preposições, pronomes e artigos. Esses componentes, para o fim a que se destina a análise, não possuem valor.

A Figura 13 exemplifica o processo de remoção de *stopwords:*

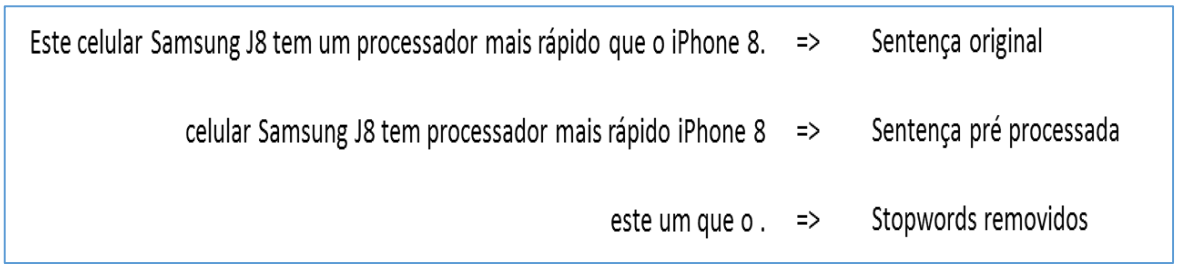

Figura 13. Etapa de remoção de *stopwords.*
Fonte: o próprio autor.

A próxima etapa foi realizar a tokenização e *part-of-speech* (POS), sendo o primeiro o processo que adiciona caracteres especiais que ajudam a delimitar as palavras (Feldman & Sanger, 2007) e o segundo comumente embasado em um dicionário de dados e tem finalidade de categorização sintática do termo. O módulo de *tagging* do NLTK, que realiza o POS, foi aplicado em seguida, classificando os elementos da sentença sintaticamente. A Figura 14 exemplifica os dois processos:

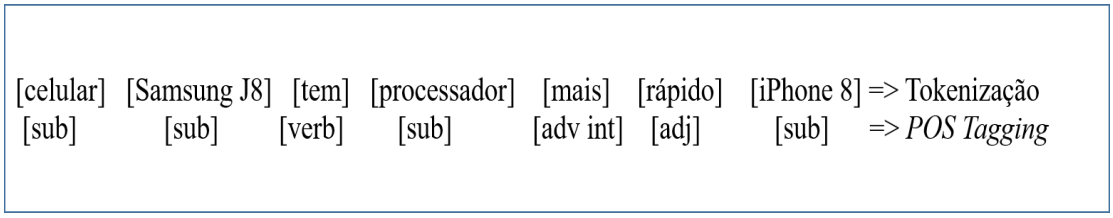

Figura 14. Etapa de tokenização e POS *Tagging.*
Fonte: o próprio autor.

Segundo Siqueira (2010), características ou aspectos são comumente substantivos, os mais frequentes no *corpus*. Isso pode ser dado, pois pessoas tendem a repetir palavras que definem uma característica. Já as demais palavras da sentença tendem a ser menos frequentes, pois são na verdade explicações relativas àquela característica.

Nessa etapa (subprocesso: transformação), foi utilizada a técnica de análise de contagem de frequência de palavras. Essa técnica foi escolhida devido à oportunidade que esse projeto possui de transformar uma desvantagem dessa abordagem em vantagem, cuja técnica não é eficiente no caso de uma amostra pequena, com poucos registros. O *corpus* desta pesquisa, porém, pode ser aumentado com pouco esforço de implementação.

Hu & Liu (2004) descrevem esse método de análise de contagem da frequência de palavras. Basicamente eles atribuem mais peso às palavras que mais aparecem, com isso são elas as consideradas características ou aspectos mais exaltados daquele produto ou serviço.

Quanto às ferramentas para análise de sentimentos, com o aumento considerável de ambientes de compartilhamento e troca de informações, como *blogs*, fóruns, mídias sociais, etc., nas mais variadas formatações textuais, ocorreu naturalmente a necessidade de implementação de métodos adequados a cada caso.

Há disponíveis, quer seja para testes ou até mesmo para compra, algumas aplicações que implementam técnicas e métodos voltados para determinados ambientes, *como SenticNet, Linguistic Inquiry and Word Count* (LIWC)*, SentiStrength, SentiWordNet*, *Sail Ail Sentiment Analyser* (SASA), etc.

O LIWC trata-se de um *software* destinado à análise de texto e que calcula o grau de utilização para diferentes categorias de palavras em uma grande variedade de textos. Ele estima componentes emocionais, cognitivos e estruturais do texto baseado em um dicionário contendo palavras e suas categorias.

O *SentiStrenght* baseia-se em aprendizado de máquina e compara métodos de classificação supervisionada e não supervisionada. Ele apresenta seus resultados em um formato de faixa em que aquele sentimento detectado é definido, como, por exemplo, não negativo (-1), extremamente negativo (-5), não positivo (1) e extremamente positivo (5).

O *SentiWordNet* é baseado no dicionário *WordNet*, que dicionário agrupa adjetivos, verbos e outras classes gramaticais. Ele associa a cada agrupamento do *WordNet* três classificações: positivo, negativo e neutro.

O *SenticNet* é baseado na análise em nível de conceito, ou seja, detecção de polaridade e reconhecimento de emoções aproveitando a semântica e linguística, sem depender exclusivamente da frequência e coocorrência de palavras. Consiste em um conjunto de técnicas de combinação de senso comum, Linguística, Psicologia e aprendizado de máquina (Sentic.net, 2016).

Já o SASA também se baseia em aprendizado de máquina, especificamente dedicada para análise de domínio do *Twitter*. Ela classifica os *tweets* como positivos, negativos, neutros ou indefinidos.

Esses aplicativos, ou comumente chamados de métodos, foram e ainda estão sendo desenvolvidos e melhorados. Inicialmente, esaa área de pesquisa concentrou

seus esforços na sumarização automatizada de informações de domínios relativos a revisões de produtos e serviços (Hu & Liu, 2004). Posteriormente, e influenciado pelo massivo uso de mídias sociais e *sites* de notícias, novas técnicas foram sendo desenvolvidas, como visto na Figura 15. Esse é o caso do *TweetSentiments* e *Sentimonitor*.

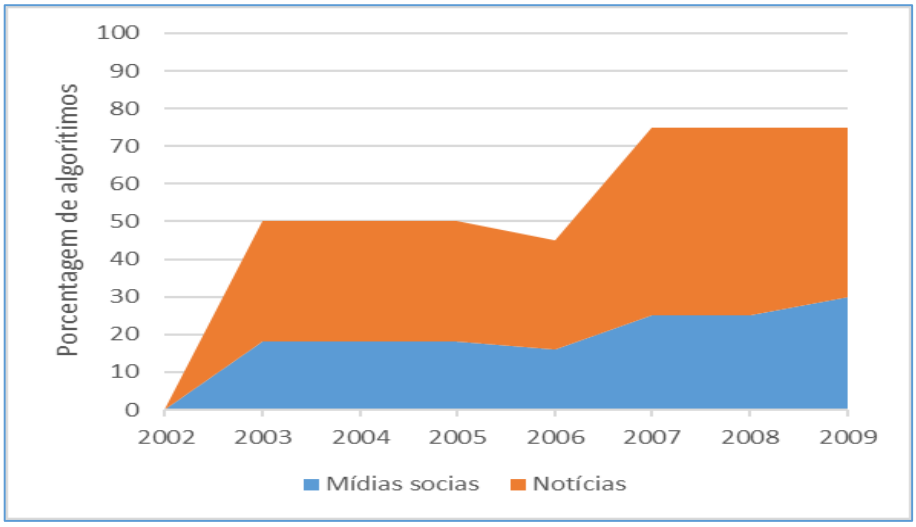

Figura 15. Evolução das pesquisas nos domínios textuais.
Fonte: adaptado de Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery, 24*(3), 478-514.

Entretanto, há características específicas de cada ferramenta, desde o idioma em que se permite realizar a extração da informação, a base textual em que se pretende avaliar (ou domínio) até mesmo métodos que implementam técnicas de inteligência artificial.

Neste trabalho usou-se o *framework* NLTK. A escolha dessa solução se dá pela relativa simplicidade de uso, mesmo para usuários sem conhecimentos avançados de programação, pela sua natureza acadêmica, que, inclusive, teve sua concepção dentro da Universidade da Pensilvânia e por ser ferramenta de uso gratuito.

Esse *framework* caracteriza-se pela multidisciplinaridade de suas ferramentas, em que cada etapa do PLN é tratada por um módulo. Como exemplo, o módulo de *token,* que provê classes para processamento e marcação dos elementos textuais, como frases e sentenças. O módulo *tree* provê mecanismo para montagem da hierarquia que define a estrutura da sentença sintaticamente. Há também o módulo *probability*, que fornece técnicas para aplicação de distribuição de frequência, probabilidade e demais técnicas estatísticas (Bird & Loper, 2004).

Na próxima etapa (subprocesso: indexação e mineração) são indexados os dados que, no contexto deste trabalho, são remetidos à ordenação de forma a propiciar acesso ágil a uma informação por meio de uma palavra-chave (Aranha & Passos, 2006). Além de facilitar esse processo de busca, permite mais de uma forma de consulta, como o uso de operadores lógicos que possam ser inseridos na *string* de busca, criando, assim, uma possibilidade de exclusão de dados indesejáveis. A Figura 16 ilustra um possível resultado dessa estrutura indexada:

Entidade	Termo	Tipo	Polaridade
celular	Samsung J8	subs	neutro
	Tem	verb	neutro
	Processador	subs	neutro
	Rápido	adj	positivo
	Mais	adv int	positivo
	iPhone 8	subs	neutro

Figura 16. Exemplo de indexação de dados.
Fonte: o próprio autor.

3.3.3 Pessoas

Conforme citado por Gil (2008), pesquisa descritiva tem como atividade-fim descrever as características encontradas em determinada população amostral, visando dar visibilidade a algum padrão descoberto entre aquelas variáveis. Dessa forma, pode acrescentar a este trabalho a projeção que se busca fazer após o devido tratamento das opiniões, em que será aplicado o PLN, e que essas projeções descobertas possam gerar diferencial competitivo àquela empresa.

Nessa etapa (subprocesso: análise), considerando-se que não há certeza de que algum padrão de fato fosse descoberto, o que foi inferido no projeto é que informações no formato da Figura 17 podem ser apresentadas nos resultados da pesquisa.

Os números, classificações e inferências projetadas, exibidos na Figura 17, foram preenchidos apenas com o efeito de ilustrar o resultado que se espera alcançar com esse experimento, não possuindo qualquer valor, portanto, não refletem a realidade.

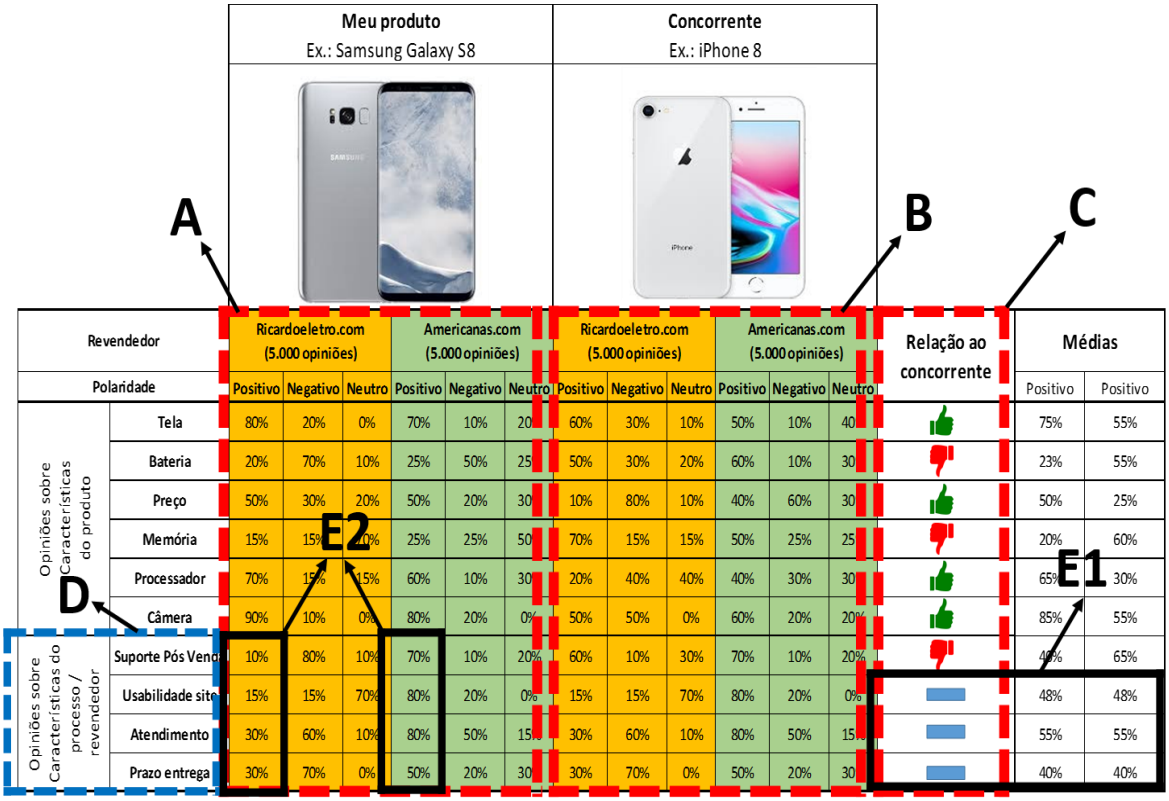

Figura 17. Projeções esperadas.
Fonte: o próprio autor.

Ana Figura 17, as colunas A e B indicam as polaridades de cada aspecto ou característica descoberta, nesse caso, sobre celulares de dois fabricantes distintos, de cada *site* revendedor do produto e de onde foram extraídas as opiniões por meio da técnica de *crawling*.

Esses aspectos ou características são os atributos do produto ou, no caso de um celular, sua tela, bateria, câmera, etc.

A coluna C indica a relação de cada aspecto tido como positivo, definido pelo processamento que será feito usando-se as técnicas de PLN explicadas neste estudo em relação a produto do concorrente, podendo ser melhor (polegar para cima), pior (polegar para baixo) e igual (traço).

Esse seria o primeiro *input* à aplicação da IC, pois, com base em uma massa amostral primária, em grande escala, aberta e coletada de maneira ética, há condições de indicar em que aspecto nosso produto tem vantagem em relação ao produto concorrente e em qual aspecto é preciso melhorar.

Pretendeu-se, também, destacar aspectos não pertencentes ao produto, como é o caso dos itens indicados pela letra D. Nesse caso, são chamados, a partir de

agora, de aspectos orbitantes, pois, apesar de não pertencerem ao produto, exercem influência indireta, como é o caso de suporte pós-venda, prazo de entrega, etc.

Caso esses aspectos orbitantes possam de fato ser recuperados e analisados, haverá um *input* de grande potencial para IC e uma oportunidade de melhoria, pois significa não uma necessidade de ajuste de portfólio ou aspecto do meu produto, mas de ajuste de processo, nesse caso do processo de revenda. Pode-se inferir, por exemplo, que um produto não precisa ser modificado para vender mais, mas que melhorar a qualidade dos serviços do revendedor, como na redução do prazo de entrega e qualidade do suporte pós-venda, venha fazer o cliente se fidelizar mais e, por consequência, postar mais opiniões positivas que possam influenciar mais pessoas a comprar aquele produto e não o do concorrente.

Outro exemplo de projeção pode ser o indicado pelas letras E1 e E2. Percebendo-se em E1 que três aspectos orbitantes têm a mesma relação de positividade do concorrente, ou seja, a princípio nenhuma necessidade de melhoria, se analisados os dados de E2 isoladamente pode-se observar que um dos revendedores não tem boa reputação naqueles aspectos orbitantes em relação ao revendedor concorrente também do mesmo produto.

Mais uma vez um *input* importante para IC pode ser descoberto, mostrando que aquele revendedor precisa ser mais bem desenvolvido para que a percepção do produto possa ser positiva quando adquirida por aquele canal de distribuição especificamente.

No tocante ao que essas projeções possam servir de insumo à IC, acredita-se que ela “envolve a identificação mais cedo possível de potenciais riscos e oportunidades, pelo agrupamento e análise de informações sobre o ambiente para suportar os gerentes a tomar decisões estratégicas para empresa” (Xu *et al*., 2011, p. 1).

4 Resultados Encontrados

Os dados obtidos e que compuseram o *corpus* deste trabalho foram obtidos no *site* da *www.amazon.com.* Foram escolhidos quatro produtos distintos, de dois fabricantes de celular atualmente em atividade.

Os produtos escolhidos foram:

a) Samsung Galaxy S8 64GB – desbloqueado
b) Samsung S5 16GB – desbloqueado
c) *iPhone* 7 32GB – desbloqueado
d) *iPhone* 5S 16GB - desbloqueado

A escolha dos produtos levou em consideração requisitos de preço, em que se buscaram produtos de fabricantes distintos, mas com preços o mais similarmente parecidos, requisitos de data de lançamento do produto no mercado. Foi escolhido um produto mais recente, mas não o mais recentemente lançado (o que poderia prejudicar a coleta de amostras) e um mais antigo e financeiramente mais acessível, porém que ainda estivesse com sua oferta de venda à disposição do cliente.

O processo de coleta gerou 1.055 opiniões sobre os quatro celulares escolhidos. Foi considerado o intervalo de data da mais recente postagem até no máximo 12 meses no passado. Essas opiniões foram armazenadas em arquivos e posteriormente tratadas, seguindo o fluxo de etapas previamente estabelecido.

O resultado desse processamento, após remoção de *stopwords*, tokenização e *POS-Tagging*, foi uma lista de palavras classificadas como aspectos por meio do módulo de Extração de Aspectos do NLTK, que usa como dicionário léxico o *WordNet*. Essa é uma das mais importantes bases de dados sobre a língua inglesa.

A seguir, pequena amostra dos tipos de aspectos detectados.

ANDROID	VERIZON	CAMERA	PRICE
APPS	SOFTWARE	PURCHASE	ISSUE
BATTERY	DISPLAY	S5	BOX
BATTERY LIFE	PRODUCT	PROBLEM	APPLE
BIXBY	USAGE	UPDATE	IPHONE
BUTTON	LIFE	MONEY	VERIZON
CAMERA	SOMETHING	CARRIER	SELLER
CASE	DAY	REASON	APPLE STORE
FEATURE	ISSUE	SIM CARD	EXPERIENCE
GLASS	REVIEW	GOOGLE	CONDITION
LOT	SETTING	APP	SHUTTER SOUND
PHONE	IPHONE	SCREEN	MESSAGE
S8	PLENTY	NOTE	CABLE
SAMSUNG	OPTION	STUFF	MANUFACTURER
SCREEN	POCKET	NOTHING	PACKAGING
THING	COMPANY	SIZE	PRICE
TIME	WEEK	MINE	CONDITION
WIFI	BRIGHTNESS	THING	WARRANTY

O módulo do NLTK que norteia a orientação das palavras definidas como aspectos anteriormente também faz uso do *WordNet* e é conhecido como *Opinion Words*. São frequentemente adjetivos, verbos e advérbios e às vezes a combinação destes dois últimos.

Por fim, após a fase de indexação dos dados coletados, foram agrupados por grupo 1 e 2 e grupo 3 e 4, sendo:

a) Grupos 1 e 2: celulares de empresas distintas, recentemente lançados e de preço mais elevado;
b) grupos 3 e 4: celulares de empresas distintas, com data de lançamento mais antiga, com preços mais acessíveis e próximos e ainda disponíveis para venda nos *sites* de compra.

Na indexação foram considerados os 15 aspectos mais relevantes de cada um dos grupos. Essa redução a apenas 15 itens foi devida à relação extensiva de aspectos mapeados, mas, que por vezes, indicou uma frequência insignificante de ocorrência do termo na amostra, tornando-se desprezível nesta avaliação. Em média, 65% da amostra tiveram aspectos citados apenas entre duas e quatro vezes. Foi desprezado também o aspecto de maior citação em todos os quatro grupos

escolhidos, nesse caso o termo "*phone",* pois se trata de termo genérico considerando que está sendo tratado o produto celular, portanto, sem valor a análise.

A coluna "citações" lista os dados classificados de forma decrescente, desde a maior frequência de citações daquele termo na amostra analisada até a menor. Também foram apresentadas as orientações de polaridade de cada aspecto, podendo ser positivo, negativo e neutro, que foram previamente calculadas pelo módulo de tratamento de opiniões do NLTK. Este aplica peso ao termo com base nas palavras de opinião detectadas e classificadas com base no dicionário *WordNet*. Os Quadros 2 e 3 demonstram o resultado da indexação feita, sendo que no Quadro 2 encontra-se o agrupamento dos produtos de maior valor e mais recentemente lançados.

Quadro 2

Indexação dos grupos 1 e 2

Grupo 1						**Grupo 2**				
Samsung Galaxy S8						**Apple iPhone 7**				
Aspectos	**Citações**	**Positivo**	**Negativo**	**Neutro**		**Aspectos**	**Citações**	**Positivo**	**Negativo**	**Neutro**
SCREEN	21	21.12	11.18	67.7		SELLER	23	27.27	16.36	56.36
S8	11	34.21	10.53	55.26		IPHONE	16	28.81	16.95	54.24
CASE	11	23.81	30.95	45.24		TIME	13	23.21	17.86	58.93
THING	9	26.47	10.29	63.24		APPLE STORE	12	11.11	18.52	70.37
SAMSUNG	9	13.46	13.46	73.08		VERIZON	10	29.63	22.22	48.15
GLASS	8	20.29	23.19	56.52		BOX	7	15.0	15.0	70.0
WIFI	8	25.0	30.0	45.0		CONDITION	7	28.21	15.38	56.41
BATTERY	7	30.65	9.68	59.68		THING	6	26.47	22.06	51.47
BATTERY LIFE	7	31.43	11.43	57.14		SCREEN	6	30.0	25.0	45.0
BIXBY	7	23.08	34.62	42.31		CAMERA	6	38.1	4.76	57.14
TIME	7	26.67	13.33	60.0		EXPERIENCE	5	26.92	15.38	57.69
BUTTON	7	18.52	25.93	55.56		PRICE	5	42.11	31.58	26.32
APPS	7	12.82	15.38	71.79		MINE	5	23.08	23.08	53.85
LOT	6	21.88	25.0	53.12		APPLE	5	24.53	13.21	62.26
% BATTERY	3	33.33	22.22	44.44		ISSUE	5	16.13	35.48	48.39

Fonte: o próprio autor.

Já o Quadro 3 ressalta os produtos de mais baixo custo e com data de lançamento mais antiga, porém ainda disponíveis para venda. Infere-se que, por serem produtos mais acessíveis, haverá também opiniões de um grupo de consumidores diferente daquele dos produtos dos grupos 1 e 2.

Quadro 3

Indexação dos grupos 3 e 4

Grupo 3						Grupo 4				
Samsung S5						Apple iPhone 5S				
Aspectos	Citações	Positivo	Negativo	Neutro		Aspectos	Citações	Positivo	Negativo	Neutro
BATTERY	18	30.77	23.08	46.15		BATTERY	22	17.86	32.48	51.19
CHARGER	14	13.64	13.64	72.73		CONDITION	13	30.43	19.57	50.0
SELLER	10	37.04	14.81	48.15		TIME	12	22.03	11.86	66.1
AMAZON	10	17.02	29.79	53.19		IPHONE	11	46.15	7.69	46.15
TIME	8	11.76	5.88	82.35		DAY	8	19.61	11.76	68.63
SAMSUNG	8	19.05	23.81	57.14		SHUTTER SOUND	7	36.11	44.44	19.44
[REVENDEDOR]	6	16.0	40.0	44.0		THING	7	18.92	16.22	64.86
CARD	6	22.22	24.07	53.7		CHARGE	7	33.33	20.83	45.83
DAY	6	29.82	24.56	45.61		ORDER	5	19.05	16.67	64.29
MONEY	6	16.67	16.67	66.67		PRICE	4	22.22	22.22	55.56
REFUND	4	20.0	30.0	50.0		JAPAN	4	30.77	26.92	42.31
PIECE	4	25.0	8.33	66.67		USE	4	18.64	8.47	72.88
SOMEONE	3	27.27	18.18	54.55		PURCHASE	4	34.78	17.39	47.83
THING	3	26.09	21.74	52.17		CAMERA SHUTTER	4	32.14	46.43	21.43
SELLER [REVENDEDOR]	3	20.0	40.0	40.0		CAMERA NOISE	3	20.0	40.0	40.0

Fonte: o próprio autor.

De posse desses dados, a atividade de gerar projeções textuais, previamente explicada no item 3.3.3, no subprocesso “análise, da fase pessoas”, pôde ser realizada.

Essa projeção textual pode ser mais bem definida como sendo inferências feitas com base em análise dos dados tratados, com informações quantitativas e qualitativas sobre a amostra, como frequência de citação e polaridade do termo. E visa identificar oportunidades de melhorias no produto ou processo, portanto, auxiliando a empresa a projetar alternativas estratégicas que lhe tragam algum tipo de vantagem competitiva.

Como se trata de pesquisa descritiva, que visa descrever características de determinada amostra a fim de obter possíveis correlações, buscou-se primeiramente identificar aspectos em comum entre os produtos, considerando apenas os 15 mais citados. A indicação de neutralidade dos aspectos será desconsiderada nesta análise, servindo apenas como dado histórico.

O aspecto “*battery”* foi citado em três grupos (1, 3 e 4), tendo positivamente maior classificação para dois produtos de um mesmo fabricante, como visto nos grupos 1 e 3. Já sua amostra para o grupo 4 obteve classificação mais negativa. Considerando os dois produtos mais bem classificados em relação ao termo *battery*, um mais recente (grupo 1) e outro mais antigo (grupo 3), observa-se que os produtos dessa empresa têm mantido qualidade ao longo do tempo em relação ao aspecto

bateria, sendo, portanto, positivamente classificado pelos consumidores ao longo do tempo.

Segue-se pequena amostra de opiniões dos grupos 1 e 3, que citaram o aspecto *battery*, provenientes do *corpus* original, sem algum tratamento:

a) *"So far, it works great. The battery typically lasts one to two days of moderate use without charging and charges quickly with the provided charger."*
b) *"Power saving/Battery life. Simply put, the AMOLED allows for some amazing battery life and the 3000mah battery holds up well. Very well. Much better than the iPhone."*
c) *"Battery Capacity. Holy cow. This thing actually lasts me over 10 hours of almost constant Netflix streaming with brightness at 75% with bluetooth wireless earbuds and that only bring its to about 30-35% battery. It then still lasts the rest of the day at 25% brightness taking it down to just about 15% battery. (I normally start charging it up once it hits 20% though)."*
d) *"This phone was able to hit all the things that I want in a phone: Good camera, long battery life, adjustable storage, crisp display, and functional."*
e) *"Battery life is very good, also charges very quickly with the included Fast Charger."*

Também foi possível identificar um padrão recorrente quanto a determinado aspecto, nesse caso, o aspecto "*shutter*" (obturador da lente da câmera):

Aspectos	Citações	Positivo	Negativo	Neutro
SHUTTER SOUND	7	36.11	44.44	19.44
CAMERA SHUTTER	4	32.14	46.43	21.43
CAMERA NOISE	3	20.0	40.0	40.0

Eles foram citados na relação de aspectos do grupo 4 em três situações, sendo: *shutter sound, camera shutter* e *camera noise.* A relação de polaridade foi negativamente mais classificada em todas as citações em que esses aspectos foram identificados.

Buscou-se, no *corpus* inicial (nos dados primários, sem algum tipo de tratamento), pelos três termos citados e foi possível abstrair que o produto do grupo 4, versão japonesa, possui uma configuração que impossibilita a desativação ou

alternativa de trocar o som típico durante o momento de tirar uma foto, que imita o som de um obturador de máquinas de fotografia. Este é um item de crítica constante.

Percebeu-se, portanto, que há uma oportunidade aqui de melhoria tanto no produto quanto no processo. No caso do produto, sugere-se a disponibilização de uma funcionalidade que permita a mudança ou desativação do som, se o celular for usado em um país onde não há legislação contrária (como declarado na seguinte opinião também contida no *corpus* original: "*Make sure you find out what country youR phone was intended for, when manufactured!! The one I purchase was originally made for Japan, where it is illegal to have the sound of the camera shutter turned off"*).

Quanto à melhoria de processo, a informação de restrição relativa a essa funcionalidade deve ser bem explicitada no momento da compra, contribuindo, assim, para possível aumento da satisfação do cliente com o produto ou, pelo menos, evitando surpresa e consequentes opiniões negativas por falta de conhecimento da restrição.

A seguir é apresentada pequena amostra de opiniões que citaram os aspectos "*shutter sound", "camera shutter"* ou *"camera noise*", provenientes do corpus original e sem algum tratamento:

a) *"The camera shutter sound would not shut off, no matter what I did. "*
b) *"Camera shutter sound never turns off, phone will die if there is anything less then 15%."*
c) *"The phone works well enough except for the fact that apparently it was originally from Japan and, as such, it is impossible to turn off the shutter sound. "*
d) *"No way to silence photo shutter noise. "*
e) *"The Japanese model in which you cannot turn off the shutter sound. "*
f) *"Only one thing: it's from Japan or somewhere where they aren't allowed to silence the camera noise, so the shutter noise goes off at the loudest volume EVERY SINGLE TIME you take a picture. "*
g) *"Also, along with all other unlocked iPhones, it had a loudass camera shutter sound that you couldn't turn off. "*
h) *"Can't seem to figure out why the shutter sound on the camera and Snapchat won't silence. "*

Uma das proposições deste trabalho foi a possibilidade de identificação de aspectos previamente classificados como orbitantes, aqueles não relacionados diretamente ao produto, mas que podem influenciar a percepção do cliente final, positiva ou negativamente. Foi possível observar dois aspectos orbitantes citados nos dados do grupo 3, que após consulta no *corpus* original, pôde-se constatar se tratar de um revendedor que opera com a venda de produtos *refurbished*, ou seja, produtos recondicionados.

Será suprimido o nome da empresa, sendo identificado a partir deste momento apenas como [REVENDEDOR].

Aspectos	Citações	Positivo	Negativo	Neutro
[REVENDEDOR]	6	16.0	40.0	44.0
SELLER [REVENDEDOR]	3	20.0	40.0	40.0

Apurou-se que os clientes que citaram esses aspectos compraram direto desse revendedor, mas que a entrega foi feita por outra empresa, por isso as opiniões foram feitas no *site* da empresa que fez a intermediação. Entretanto, reclamavam que produtos com defeito foram enviados, como cartões de memória que não eram reconhecidos pelo celular, além de citarem que produtos usados eram enviados classificados como novos.

Aqui pode ser detectada uma oportunidade de restrição de venda do produto do grupo 3 por esse revendedor ou possível desenvolvimento de um *checklist* de qualidade pré-venda mais assertivo, para garantir a checagem dos produtos encomendados antes do envio ao cliente.

Transcreve-se pequena amostra de opiniões que citaram aspectos orbitantes relativos ao revendedor, provenientes do *corpus* original do grupo 3:

a) *“Well I bought two of these phones from a seller called* [revendedor] *but they were Fulfilled by Amazon. One of them has been awful since it arrived. When it was first turned on my wife told me she thinks someone had the phone before but never made a thing of it.”*
b) *“When I thought the nightmare was over as they resolved it by allowing me to return that phone, we now find out two days later that the other one we purchased from* [revendedor] *can't read memory cards. I purchased two new*

memory cards for a new phone we ordered which was the same Samsung Galaxy s5 as we like the phone and have concluded that the seller [revendedor] *is very sketchy and unscrupulous."*

c) *"Seems a review from someone saying this seller is sending out phones as new that were already used and had issues but I was not sure because the review had seemed like it was more about the product and was not 100% sure it was that seller* [revendedor]*."*

Alinhado ao problema de pesquisa desta dissertação - gerar projeções textuais no sentido de antecipação de situação futura ou determinação de preferências -, pôde se projetar nos resultados encontrados oportunidades de melhoria para o produto do grupo 4, no aspecto de autonomia da carga da bateria. Os produtos equivalentes de seu concorrente aqui em análise, nesse caso os produtos dos grupos 1 e 3, têm esse aspecto recorrentemente sendo bem avaliado pelos consumidores ao longo dos anos, como pode ser identificado nos dois celulares desse fornecedor, sendo um de lançamento mais recente e outro mais antigo.

Considerando que tempo é um item de extremo valor, é natural notar que as pessoas optam por consumir produtos que trarão vantagens quanto à superioridade do tempo de autonomia de energia. Percebe-se aqui agregação de valor significativa para o processo de IC da empresa, pois um dos produtos em análise tem aprovação reconhecida pelo consumidor e já de forma recorrente ao longo do tempo. Portanto, serve de alerta ao fabricante do produto concorrente, que precisa melhorar esse aspecto.

Resultados similares aos que aqui foram encontrados podem ser vistos nos trabalhos de Akhtar *et al*. (2017). Também foi proposta uma abordagem a partir da análise de opiniões baseadas em aspectos, cujas etapas de processamento do *corpus* podem ser vistas na Figura 18. Após o tratamento do *corpus*, este pôde servir de insumo para a inteligência competitiva, apoiando a busca por melhores hotéis nas *reviews* feita pelos consumidores nos *sites* hoteleiros.

Foi feito inicialmente um *crawling* das opiniões*,* porém neste trabalho não foram observadas melhorias em termos de automatização do método. Foi passada apenas a URL do *site* em que se deseja extrair o *corpus*, mas sem indicação de que a função de *crawling* tenha percorrido as várias páginas que continham opiniões ou se todas as opiniões foram coletadas.

Na fase de pré-processamento, a mesma abordagem foi feita como proposta nesta dissertação, portanto, usaram-se os componentes de PLN do NLTK.

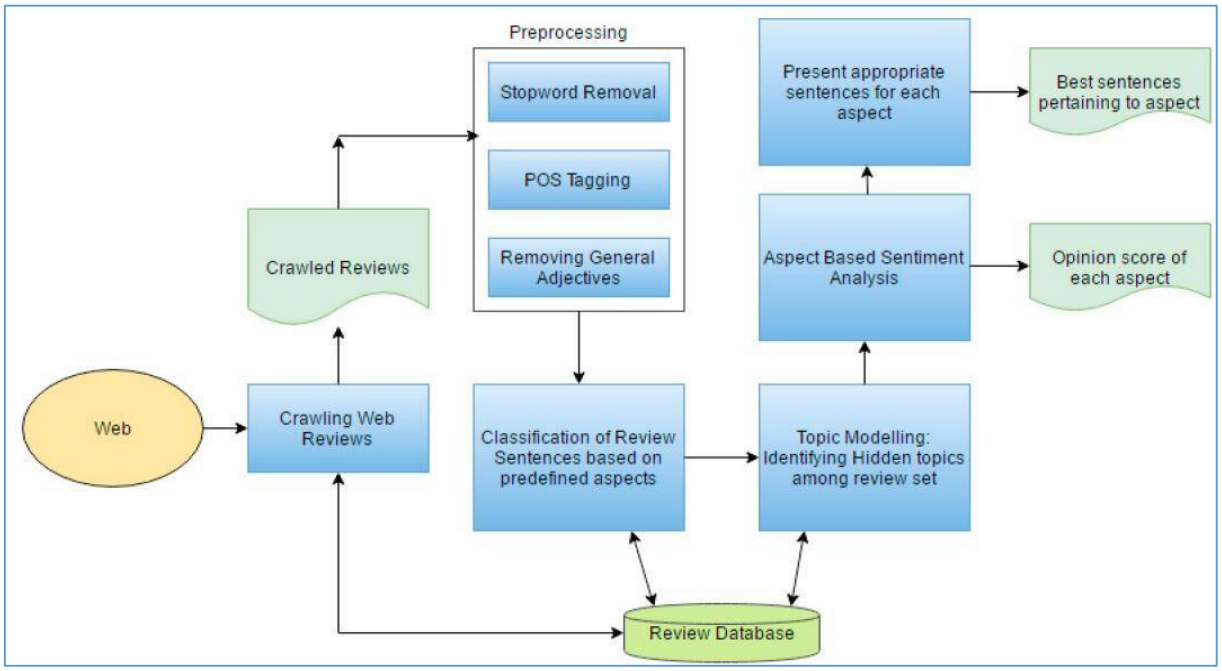

Figura 18. Processo proposto para análise de opiniões baseada em aspectos.
Fonte: Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect based sentiment oriented summarization of hotel reviews. *Procedia Computer Science, 115*, 563-571.

Os resultados encontrados revelaram relação de aspectos que foram posteriormente ordenados por segmento, como exemplo *Room* (*bed, clean, toilet, shower, etc)* e *Facility* (*pool, spa, gym)*, em seguida os aspectos tiveram sua polaridade classificada.

Essa relação de aspectos polarizados e agrupados servirá de base para que os hotéis possam aprimorar itens com pior classificação ou com mais opiniões negativas. E também servirá ao consumidor final, uma vez que pode tomar suas decisões com base em termos específicos que melhor classificariam sua busca, em vez de simplesmente ler milhares de opiniões, que muitas vezes podem ser ambíguas ou tendenciosas.

Outro trabalho que fez uso de procedimentos similares aos desta dissertação, como dito, foi o de Symeonidis *et al*. (2018). Nele o módulo do NLTK de pré-processamento foi usado como ferramenta na etapa de remoção de *stopwords,* porém demonstrou que o uso dessa ferramenta quando o *corpus* é proveniente do *Twitter* não se mostra eficaz, pois na maior parte o público dessa mídia social é composto de

pessoas jovens, que usam com muita frequência termos curtos e com muitas gírias, gerando, assim, falso-positivos como se fossem de fato *stopwords*.

Outro trabalho que fez uso do mesmo domínio desta dissertação - celulares - e que também foi extraído do *site* da *Amazon.com* é o de Chawla, Dubey & Rana (2017). Nele foram analisadas *reviews* sobre celulares, com a finalidade de comparar duas abordagens para detecção de polaridade, uma por meio do algoritmo *Naïve Bayes* e outra por meio do *Support Vector Machine* (SVM).

A motivação foi apresentar qual das ferramentas oferece mais precisão na classificação das opiniões, por se entender que existe cada vez mais necessidade de as empresas saberem o que seus consumidores estão efetivamente pensando e divulgando sobre seus produtos. Isso porque mais e mais mídias sociais estão acessíveis e são um ambiente de formação de opinião muito abrangente.

Os autores coletaram 1.000 postagens por meio de técnica de *crawling,* sem automatismo algum no sentido de percorrer todas as páginas de opiniões, que, por sua vez, geraram aproximadamente 3.000 sentenças. Essas opiniões são relativas aos *smartphones* Samsung, Micromax e One Plus, mais precisamente sobre os aspectos a eles associados.

Após as etapas de pré-processamento do *corpus*, no qual foram aplicadas as técnicas de remoção de *stopwords* e *Part-of-Speech*, foi possível apresentar os aspectos mais citados, exemplificados por meio do conceito de nuvem de palavras, conforme Figura 19.

Figura 19. Nuvem de palavras dos aspectos detectados.
Fonte: Chawla, S., Dubey, G., & Rana, A. (2017). Product opinion mining using sentiment analysis on smartphone reviews. *6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (ICRITO), India.

Esse conjunto de dados tratados serviu de entrada para a aplicação do algoritmo *Naïve Bayes* e do SVM para a comparação da precisão da classificação. O resultado da pesquisa demonstrou resultado de 40% de precisão no uso do *Naïve Bayes* e 90% no uso do SVM, sendo este segundo, portanto, a recomendação quando a análise da polaridade de *reviews* for requerida, independentemente do *corpus* tratado.

5 Considerações Finais

Considerando que o processo de IC "objetiva agregar valor à informação, fortalecendo seu caráter estratégico, catalisando, assim, o processo de crescimento organizacional" (Canongia, 1998, p. 2-3), a abordagem proposta neste trabalho por meio do uso de PLN trouxe evidente vantagem estratégica. Tratou e gerou informações com valor agregado que, quando utilizadas pelas empresas, poderão contribuir no processo de melhorias dos produtos e no processo de venda. Além de detectar necessidade de melhoria em um aspecto inerente ao produto (bateria), acusou também um aspecto orbitante que diz respeito ao processo de venda do produto. Nesse segundo caso, uma oportunidade de desenvolvimento do revendedor poderia ser alternativa para melhoria da percepção do consumidor.

A dissertação contribuiu também para algumas das principais atividades do processo de mineração de opiniões, aprimorando o processo de extração do *corpus* por meio de adição de funcionalidades à biblioteca jSoup, melhoria ao código, aumentando o grau de automatização da extração dos dados. Essa etapa é tida como onerosa, exatamente pela falta de automatismo adequado, consumindo, dessa forma, muito tempo e não garantindo a extração de suficiente amostra.

Pôde-se observar o valor desse automatismo da fase de coleta de dados, quando confrontado com os resultados do trabalho de Stieglitz, Mirbabaie, Ross & Neuberger (2018), que salientaram os desafios relativos ao processo de análise de mídias sociais, que tipicamente se divide em quatro fases: descoberta, coleta, preparação e análise de dados. Representativo volume de trabalhos já existe para a fase de análise de dados, porém não existe na mesma proporção para as demais fases, especificamente para a descoberta e coleta.

A pesquisa desses autores foi feita em cinco diferentes bases de conhecimento, sendo a *string* de busca apresentada na Tabela 2. O resultado retornou 260 artigos, dos quais 49 foram relevantes. Destes documentos relevantes, foi possível observar que, nas três primeiras fases, o maior desafio encontrado relacionou-se ao volume de dados disponíveis, que por vezes se mostrou sem a qualidade ou aderência adequadas ao que se buscava. Portanto, o excesso ou falta de dados pode não ser suficientemente representativo como amostra, fazendo com que um mecanismo que facilite o volume de dados buscados se torne um elemento de real valor ao processo.

Tabela 2

String de busca e bases de conhecimento usadas

Search terms			Databases	Fields
("social media analytics"	AND	(challenges	ACM Digital Library	Title, Abstract
OR "social media analysis"		OR difficulties	AIS Electronic Library	
OR "social media data"		OR problems	IEEE Xplore	
OR "social media mining")		OR complexity)	ScienceDirect	

Nota. Fonte: Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156-168.

Na continuidade do presente estudo, foram feitos a identificação dos aspectos-alvo das opiniões, seu pré-processamento e posterior classificação da polaridade desses aspectos, fornecendo como resultado final tabelas com dados indexados com elementos quantitativos e qualitativos, que puderam ser analisados visando à busca por padrões que norteassem a construções de projeções de melhorias do produto ou processo.

O resultado mostrou-se efetivo no sentido de realçar como o uso de ferramentas de PLN de forma estruturada pode ser valioso para o tratamento de dados, além de agregar-lhes valor. Além disso, contribuiu com o detalhamento do passo a passo para uso de componentes prontos, como é o caso dos módulos do NLTK, que possuem média complexidade no sentido de interação do usuário final, podendo, portanto, ser replicado por pessoas com algum conhecimento em programação e lógica.

As projeções apresentadas no capítulo 4, baseadas nos dados dos Quadros 2 e 3, foram feitas por uma abordagem descritiva, conforme previsto, onde se buscou algum tipo de correlação de variáveis e padrões de comportamento. E isso se deu para poder exemplificar como este trabalho pode contribuir em relação ao resultado apresentado, pois teve como meta fornecer à empresa dados suficientemente tratados e com algum valor agregado, em contrapartida ao dado anteriormente em sua forma original, apenas em formato não estruturado, com considerável uso de termos com pouco valor informacional ao que se propõe e com possíveis tendências ou até mesmo incoerência no discurso, situação típica de quando se expressa de forma textual.

A geração dessas projeções por meio da busca por padrões nos resultados e correlação de variáveis se tornará mais assertiva e efetivamente mais produtividade, se os dados puderem ser analisados por um profissional da própria empresa, ou seja, alguém que possua entendimento e orientação sobre qual informação tem real valor e que esteja alinhado às estratégias e aos interesses da empresa. Dessa forma, a aplicação das etapas propostas neste estudo para tratamento das opiniões pode se tornar efetivamente um elemento de diferencial competitivo para aquela empresa.

As projeções descritas foram geradas e inferidas pelo próprio autor da dissertação, que não tem ligação com as empresas citadas, tampouco possuía conhecimento prévio sobre a manufatura de celulares e seu ecossistema (vendas, suporte, garantia, etc.) para decidir se de fato as informações encontradas têm efetivo valor estratégico. Mas serviu para demonstrar que mesmo com restrito conhecimento sobre o tema, é possível até mesmo ao leigo buscar conhecimento que possa ter valor e orientar alguma tomada de decisão futura ou ajuste imediato no portfólio em questão.

5.1 Limitações da pesquisa

Algumas das limitações deste trabalho serão relatadas a seguir. São esses elementos que permitirão a sua sequência em um novo projeto de pesquisa:

a) Como a intenção do experimento foi demonstrar o uso de técnicas de PLN, contribuindo efetivamente para o processo de IC, os dados não foram tratados e classificados por outras ferramentas que não as do pacote NLTK;
b) os resultados não foram comparados com os de outras abordagens de PLN, portanto, é de se esperar que em trabalhos futuros o mesmo *corpus* possa sofrer alterações em sua classificação final, caso sejam submetidos a outras ferramentas;
c) como informado anteriormente, os dados da Figura 17 foram inferidos pelo autor da dissertação, servindo exclusivamente como ilustradores do resultado que se esperava alcançar com este experimento. Logo, não possuem algum valor real e não refletem a realidade.

5.2 Trabalhos futuros

Cada uma das etapas aqui apresentadas fez uso de componentes de *software* de forma modular, como foi o caso da fase de *crawling* usando uma biblioteca Java e a fase de pré-processamento, que fez uso dos componentes do NLTK.

Como proposta de trabalho futuro, existe a oportunidade de construção de um aplicativo que agrupe em uma só interface todas as etapas citadas anteriormente. E que estas sejam transparentes ao usuário, ou seja, que a única interação com essa aplicação se dê pelo usuário ao definir determinado produto, seu produto concorrente e a fonte de opiniões sobre eles. Como resultado, que sejam apresentados os dados já indexados, com os aspectos mais relevantes descobertos, sua polaridade e alguns gráficos, que poderão subsidiar a construção das projeções, sem necessidade de conhecimento prévio de linguagens de programação, para instalação dos componentes de *software*, como o Java e o NLTK, a preparação do *corpus*, as fases de pré-processamento e o tratamento para indexação dos resultados.

Outra possibilidade de trabalho futuro é a aplicação do *corpus* trabalhado neste experimento a outros métodos de PLN, visando uma comparação entre as abordagens. Dessa forma, poderia ser comparada a precisão de cada abordagem.

Referências

Akhtar, N., Zubair, N., Kumar, A., & Ahmad, T. (2017). Aspect based sentiment oriented summarization of hotel reviews. *Procedia Computer Science, 115*, 563-571.

Amarouche, K., Benbrahim, H., & Kassou, I. (2015). Product opinion mining for competitive intelligence. *Procedia Computer Science, 73*, 358-365.

Andrade, F. W. G. (2018). *Mineração de texto: previsão de tendências no mercado de ações por meio de notícias publicadas na internet*. Dissertação (Mestrado em Administração) - Fundação Mineira de Educação e Cultura – FUMEC.

Aranha, C. N., & Vellasco, M. (2007). *Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional*. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, RJ.

Aranha, C., & Passos, E. (2006). A tecnologia de mineração de textos. *Revista Eletrônica de Sistemas de Informação*. ISSN 1677-3071. Doi: 10.21529/RESI, 5(2).

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Retrieval evaluation. (Cap. 3). *In*: _____. *Modern information retrieval*. England: ACM Press.

Barros, A. J. S., & Lehfeld, N. A. d. S. (2000). *Fundamentos de metodologia científica: um guia para a iniciação científica*. São Paulo: Makron (2. ed. ampl.).

Berry, M. W., & Kogan, J. (Eds.). (2010). *Text mining: applications and theory*. New Jersey: John Wiley & Sons.

Bird, S., & Loper, E. (2004, July). NLTK: the natural language toolkit. *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions,* Barcelona (p. 31). Association for Computational Linguistics.

Birjali, M., Beni-Hssane, A., & Erritali, M. (2017). Machine learning and semantic sentiment analysis based algorithms for suicide sentiment prediction in social networks. *Procedia Computer Science, 113*, 65-72*.*

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8.

Canongia, C. (1998). Sistema de inteligência: uso da informação para dinamização, inovação e competitividade. *Simpósio Internacional de Propriedade Intelectual, Informação e Ética*, *1*, Florianópolis.

Chawla, S., Dubey, G., & Rana, A. (2017). Product opinion mining using sentiment analysis on smartphone reviews. *6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)* (ICRITO), India.

Chen, H. (2001). *Knowledge management systems: a text mining perspective*. Arizona: Knowledge Computing Corporation.

Chowdhury, G. G. (2003). Natural language processing. *Annual Review of Information Science and Technology, 37*(1), p. 51-89.

Coelho, G. M., Dou, H,. Quonian, L., Silva, C. H. (2006). Ensino e pesquisa no campo da inteligência competitiva no Brasil e a cooperação franco-brasileira. *Revista Hispana de La Inteligencia Competitiva - Puzzle*. España, año 6, (23), 12-19.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge: Cambridge University Press.

Gao, S., Tang, O., Wang, H., & Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management, 71*, 19-32.

Garcia, A. E. G. (2018). Inteligência competitiva: considerações sobre a prática no ambiente empresarial brasileiro. *Revista Inteligência Competitiva, 8*(1), 127-168.

Gil, A. C. (2008). *Métodos e técnicas de pesquisa social*. (6. ed.). Belo Horizonte: Atlas.

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Rio de Janeiro: Elsevier.

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(issue 7), 801-812.

Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A brief survey of text mining. *Ldv Forum*, *20*(1), 19-62.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 168-177). ACM, Washington.

Kansal, H., & Toshniwal, D. (2014). Aspect based summarization of context dependent opinion words. *Procedia Computer Science, 35,* 166-175.

Kauer, A. U. (2016). *Análise de sentimentos baseada em aspectos e atribuições de polaridade*. Dissertação (Mestrado em Administração) - Porto Alegre: Universidade Federal do Rio Grande do Sul.

Kumar, A., & Abirami, S. (2018). Aspect-based opinion ranking framework for product reviews using a Spearman's rank correlation coefficient method. *Information Sciences*, *460-461,* 23-41.

Liu, B. (2012) Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 5*(1), 1–167.

Lopes, B., De Muÿlder, C. F., & Judice, V. M. M. (2012). Inteligência competitiva e o caso de um arranjo produtivo local de eletrônica brasileiro. *Gestão & Planejamento-G&P, 12*(2).

Maia, L. C. G., & Souza, R. R. (2010). Uso de sintagmas nominais na classificação automática de documentos eletrônicos. *Perspectivas em Ciência da Informação, 15*(1), 154-172.

Mascarenhas, S. (2010). *Metodologia científica*. Pearson Brasil, 2010. ISBN 9788564574595. Recuperado de: books.google.com.br/books?id=kOZBLgEACAAJ.

Moertini, V. S., Kevin, V., & Satyadi, J. (2017). Mining opinions from big data of indonesian hotel reviews. *Journal of Theoretical & Applied Information Technology, 95*(14).

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. *In LREc, 10*.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and trends in information retrieval. *Now Publishers Inc., 2*(1-2), 1–135.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., & Androutsopoulos, I. (2015). Semeval-2015 task 12: Aspect based sentiment analysis. *Proceedings of the 9th International Workshop on Semantic Evaluation* (SemEva), 486-495.

Qiu, G., Liu, B., Bu, J., & Chen, C. (2011). Opinion word expan sion and target extraction through double propagation. *Computational linguistics, 37*(1), 9-27.

SCIP – Society of Competitive Intelligence Professionals. About SCIP. [2018]. Recuperado de: http://www.scip.org.

Siqueira, H. B. A. (2010). *WhatMatter: extração e visualização de características em opiniões sobre serviços*. Dissertação (Mestrado em Ciência da Computação)-Universidade Federal de Pernambuco.

Souza, R. R. (2005). Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics–Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management, 39*, 156-168.

Stoyanov, V., & Cardie, C. (2006, Jul.). Partially supervised coreference resolution for opinion summarization through structured rule learning. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing,* Sidney (pp. 336-344). Association for Computational Linguistics.

Sun, S., Luo, C., & Chen, J. (2017). A review of natural language processing techniques for opinion mining systems. *Information Fusion, 36*, 10-25.

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*.

Tang, D., Qin, B., & Liu, T. (2015). Learning semantic representations of users and products for document level sentiment classification. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, 2,* Beijing, China.

Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery, 24*(3), 478-514.

Turban, E., Leidner, D., Mclean, E., & Wetherbe, J. (2010). *Tecnologia da Informação para Gestão: transformando os negócios na economia digital*. São Paulo: Bookman.

Valentim, M. L. P., Lenzi, L. A. F., Cervantes, B. M. N., de Carvalho, E. L., Garcia, H. D., Catarino, M. E., & Tomaél, M. I. (2003). O processo de inteligência competitiva em organizações. *DataGramaZero*, Rio de Janeiro, *4*(3), A03-1001.

Valsan A., Sreepriya C.T., & Nitha L. (2017). Social media sentiment polarity analysis: a novel approach to promote business performance and consumer decision-making. *In*: S. Dash, K. Vijayakumar, Panigrahi B., & S. Das (eds) Artificial intelligence and evolutionary computations in engineering systems. *Advances in Intelligent Systems and Computing, 517,* Springer, Singapore.

Vickery, T. (2018, março 23). *As bancas de revista que viraram mercados, a revolução de dados e o diagnóstico sombrio do Brasil*. Recuperado de; http://www.bbc.com/portuguese/blog-tim-vickery-43479246.

Vieira, R., & Lima, V. L. S. (2001). Linguística computacional: princípios e aplicações. *In:* L. Nedel (ed.). *IX Escola de Informática da SBC-Sul*, *SBC-Sul*.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004) Learning subjective language. *Computational Linguistics, MIT Press, 30*(3), 277–308.

Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision support systems, 50*(4), 743-754.

Yu, L., & Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research, JMLR. org, 5*, 1205–1224.

Zhai, Z., Liu, B., Xu, H., & Jia, P. (2011). Clustering product features for opinion mining. *Proceedings of the fourth ACM international conference on Web search and data mining. ACM*, Hong Kong.