UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM
ENGENHARIA E GESTÃO DO CONHECIMENTO

FLÁVIO CECI

UM MODELO BASEADO EM CASOS E ONTOLOGIA PARA APOIO À TAREFA INTENSIVA EM CONHECIMENTO DE CLASSIFICAÇÃO COM FOCO NA ANÁLISE DE SENTIMENTOS

Florianópolis
2015

FLÁVIO CECI

UM MODELO BASEADO EM CASOS E ONTOLOGIA PARA APOIO À TAREFA INTENSIVA EM CONHECIMENTO DE CLASSIFICAÇÃO COM FOCO NA ANÁLISE DE SENTIMENTOS

Tese submetida ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento da Universidade Federal de Santa Catarina para a obtenção do título de Doutor em Engenharia e Gestão do Conhecimento. Área de concentração: Engenharia do Conhecimento. Linha de pesquisa: Teoria e prática em Engenharia do Conhecimento

Orientador: Alexandre Leopoldo Gonçalves, Dr.
Coorientador: Roberto Carlos dos Santos Pacheco, Dr.

Florianópolis
2015

Ficha de identificação da obra elaborada pelo autor por meio do Programa de Geração Automática da Biblioteca Universitária da UFSC

Ceci, Flávio
UM MODELO BASEADO EM CASOS E ONTOLOGIA PARA APOIO À TAREFA INTENSIVA EM CONHECIMENTO DE CLASSIFICAÇÃO COM FOCO NA ANÁLISE DE SENTIMENTOS / Flávio Ceci ; orientador, Alexandre Leopoldo Gonçalves ; coorientador, Roberto Carlos dos Santos Pacheco. - Florianópolis, SC, 2015.
211 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Inclui referências

1. Engenharia e Gestão do Conhecimento. 2. Análise de sentimento. 3. Classificação semântica. 4. Ontologia. 5. Raciocínio baseado em casos. I. Leopoldo Gonçalves, Alexandre . II. dos Santos Pacheco, Roberto Carlos . III. Universidade Federal de Santa Catarina. Programa de PósGraduação em Engenharia e Gestão do Conhecimento. IV. Título.

FLÁVIO CECI

UM MODELO BASEADO EM CASOS E ONTOLOGIA PARA APOIO À TAREFA INTENSIVA EM CONHECIMENTO DE CLASSIFICAÇÃO COM FOCO NA ANÁLISE DE SENTIMENTOS

Esta tese foi julgada adequada para a obtenção do título de Doutor em Engenharia e Gestão do Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Florianópolis, 28 de agosto de 2015.

Prof. Roberto Carlos dos Santos Pacheco, Dr.
Coordenador do Curso

Banca Examinadora:

Prof. Alexandre Leopoldo Gonçalves, Dr.
EGC-UFSC
Orientador

Prof. Ricardo Azambuja Silveira, Dr.
INE-UFSC

Prof. José Leomar Todesco, Dr.
EGC-UFSC

Prof. Eros Comunello, Dr.
MCA-Univali

Prof. João Artur de Souza, Dr.
EGC-UFSC

Profª. Rosina de O. Weber, Dra.
Drexel University

Dedico este trabalho à minha esposa Gláucia, que sempre esteve comigo em todos os momentos dessa caminhada me dando força, apoio, carinho e atenção. Eu te amo, Gau!

AGRADECIMENTOS

Para o desenvolvimento desta tese, foi necessário muito empenho e dedicação da minha parte. Contudo, o trabalho não seria possível sem a participação direta ou indireta de algumas pessoas, as quais eu gostaria de agradecer aqui neste espaço.

Primeiramente agradeço a Deus e aos meus pais, Altamiro e Margarida, que sempre estiveram ao meu lado oferecendo apoio e assistência constantes, quando não se sacrificando para garantir uma boa formação e dando condições para o meu desenvolvimento profissional e acadêmico.

À minha esposa Gláucia, que sempre teve muita paciência nos meus momentos de crise, me apoiando nas minhas decisões. Agradeço por todo amor, carinho e atenção que sempre teve comigo.

Aos meus sogros, José João e Zuete, que me acolheram em sua família, sempre me dando apoio e atenção.

Aos meus padrinhos, Rita, Sidnei e Sérgio, pelo enorme apoio que me foi dado durante toda a minha vida. Também agradeço aos meus cunhados, cunhadas e sobrinhos pela cooperação e atenção. Principalmente à minha sobrinha Clara, que me auxiliou na correção da redação deste documento para a etapa de qualificação.

Ao meu grande amigo e orientador, Professor Dr. Alexandre Leopoldo Gonçalves, que sempre teve paciência e disposição comigo, pelas incansáveis e esclarecedoras conversas e pela orientação deste trabalho. Também agradeço ao meu amigo e coorientador, Professor Dr. Roberto Carlos dos Santos Pacheco, pela sua atenção e paciência, e pela coorientação neste trabalho.

À Professora Dra. Rosina Weber, que acompanhou todo o desenvolvimento deste trabalho, sempre dando contribuições, sugestões e ensinamentos. Também por aceitar fazer parte da banca examinadora.

Aos professores do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento que aceitaram compor a banca examinadora, Dr. José Leomar Todesco, Dr. João Artur de Souza e Dr. Denilson Sell, bem como ao Dr. Ricardo Azambuja Silveira, professor da INE-UFSC e ao Professor Dr. Eros Comunello do MCA-Univali. É uma grande honra tê-los como avaliadores deste trabalho.

Ao Instituto Stela, pela confiança e pela flexibilidade de horários, o que me possibilitou participar das atividades do Programa.

Aos meus colegas de trabalho, Dr. José Leomar Todesco, Dr. Fabiano Beppler, Márcio Napoli, Dr. Denilson Sell, Roberto Fabiano Fernandes, Dr. Jean Carlo Rossa Hauck, Flavia Lumi Matuzawa,

Mateus Andriani Lohn, Edison Aquino de Meireles Neto, Cleiton Edgar Janke Duarte, Júlio Gonçalves Reinaldo, Luyane Cardoso e Maria Elisa da Silva Rosa, pelas conversas esclarecedoras e pelo apoio em geral.

À Coordenação e aos professores dos cursos de Ciência da Computação e Sistemas de Informação da Universidade do Sul de Santa Catarina, pela oportunidade de vivenciar a experiência de docência durante a concepção deste trabalho e pelo apoio, em especial à Dra. Vera Schummacher e à Dra. Maria Inés Castiñera.

A todos os excelentes professores que tive durante a minha trajetória acadêmica, os quais me ensinaram e também inspiraram a seguir pelo caminho da docência. Em especial aos Professores Dr. Alexandre Leopoldo Gonçalves, Dr. Aran Bey T. Morales, Dra. Vera R. Schummacher, Dra. Maria Inés Castiñeira, Dr. José Leomar Todesco, Dr. Denilson Sell, Dra. Andrea Valéria Steil, M. Eng. Mauro Pacheco, Dr. Francisco Antônio P. Fialho, Dr. Aquino Lauri de Espíndola, Esp. Fernanda Oviedo Bizarro, Dr. Ricardo Villarroel Dávalos e Elis Rogéria Pelegrini.

Ao Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, pela oportunidade em participar desse prestigiado curso.

Agradeço também às demais pessoas que participaram direta ou indiretamente do desenvolvimento deste trabalho. Aos amigos de Santo Amaro da Imperatriz, em especial: João Antônio Ventura Neto, Angelo Brüggemann, Fabio Fernandes, Gabriela Brüggemann, Roberta Cristina Loch, Louiza Hüntemann Garcia e Felipe Hawerroth Hoepers.

RESUMO

O uso de informações externas às organizações, presentes nas redes sociais, pode auxiliar no processo de compra de um produto por parte de um cliente a partir da leitura de revisões ou de *blogs* especializados. A classificação de texto, mais precisamente, a análise de sentimento, auxilia na definição da imagem de um produto ou na identificação do sentimento de uma sentença como positiva ou negativa. Neste trabalho propõe-se um modelo que combina ontologia de domínio com técnicas de processamento de linguagem natural para a identificação do sentimento agregado a uma determinada sentença, buscando apresentar uma explicação para tal polarização. Utiliza-se ainda o raciocínio baseado em casos para que seja possível aprender com os raciocínios (polarizações) passados, visando reutilizá-los em novas classificações. Também foram elaboradas etapas para o tratamento de negação, para a adequação do léxico de sentimento para um domínio e para a adaptação da classificação de termos ambíguos baseados em classificações passadas. Foram desenvolvidos testes em dois domínios distintos, câmeras digitais e filmes, para justificar a evolução do modelo até se chegar à proposta final. Pôde-se observar que a acurácia obtida pelo modelo é superior à obtida por abordagens estatísticas tradicionais. Esses resultados demonstram que o modelo da tese contribui para a área de análise de sentimento, tanto no nível da acurácia quanto pela possibilidade de apresentar o caminho percorrido para chegar a determinada classificação.

Palavras-chave: Análise de sentimento, Classificação semântica, Ontologia, Raciocínio baseado em casos, Árvore de sentimento.

ABSTRACT

The use of information outside organizations available in social networks such as reviews or specialized blogs can assist customers in their decisions. The text classification, more precisely sentiment analysis, assists in defining the image of a product or identifying the sense of a sentence as positive or negative. This work intends to combine domain ontology with natural language processing techniques to identify the sentiment behind judgments aiming to provide an explanation for such polarization. Also, it intends to use the Case-Based Reasoning strategy in order to learn from past reasonings (polarizations) so they can be used in new polarizations. Some steps have been developed for treatment of negation, adequacy of sentiment lexicon for a domain and adaptation of ambiguous terms classification based on past ratings. Tests were developed in two distinct areas, digital cameras and movies, to justify the model evolution until its final proposal. It was observed that the accuracy obtained by the proposed model overcomes standard statisticals approaches. These results demonstrate that the thesis model contributes to the sentiment analysis area, both as a solution that provides high levels of accuracy, as well as the possibility to present the track to achieve a particular classification.

Keywords: Sentiment analysis, Semantic classification, Ontology, Case-based reasoning, Sentiment tree.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS

CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico
EGC – Engenharia e Gestão do Conhecimento
EWGA – *Entropy Weighted Genetic Algorithm*
FCA - *Formal Concept Analysis*
GC – Gestão do Conhecimento
K-NN - *K-Nearest Neighbors*
LDA – *Latent Dirichlet Allocation*
LOOCV - *Leave One Out Cross Validation*
LSA – *Latent Semantic Analysis*
NB – Naïve Bayes
NER – *Named Entity Recognition*
PLN – Processamento de Linguagem Natural
PLSA - *Popular Topic Modeling Algorithm*
PMI – *Pointwise Mutual Information*
POS – *Part of Speech*
RBC – Raciocínio Baseado em Casos
RI – Recuperação de Informação
SDA – *Staked Denoising Autoencoders*
SOT – *Sentiment Ontology Tree*
SVD – *Singular Value Decomposition*
SVM – *Support Vector Machine*
UFSC – Universidade Federal de Santa Catarina

SUMÁRIO

1 INTRODUÇÃO

Atualmente, uma parte considerável do conhecimento organizacional está disponível na forma de documentos textuais não-estruturados como, por exemplo, em livros, manuais, relatórios, registros de reuniões, entre outros (CHAVES, 2009).

Esses documentos, quando produzidos dentro do contexto organizacional, apresentam informações para mapear o seu domínio, de modo a compor bases de conhecimento que posteriormente possam auxiliar na tomada de decisão bem como subsidiar sistemas baseados em conhecimento.

Para representar o conhecimento organizacional de modo a facilitar a sua utilização e recuperação, ontologias vêm sendo cada vez mais utilizadas (VASCONCELOS; ROCHA; KIMBLE, 2003). Segundo Studer, Benjamins e Fensel (1998), uma ontologia é a especificação explícita e formal de conceitos e relações que existem em um domínio, e que são compartilhados por uma comunidade.

É importante observar que existem muitas informações e conhecimentos da organização presentes de maneira não-estruturada e fora do seu domínio, como, por exemplo, na Internet, mais especificamente, nas mídias sociais em geral. Essas informações podem apresentar opiniões sobre produtos ou serviços oferecidos pelas organizações.

Nesse contexto, a classificação é uma tarefa intensiva em conhecimento importante e aplicável que permite a organização e a entrega de informações categorizadas por conceitos do seu domínio. Segundo Freitas (1997), a tarefa de classificação tem como função prever o valor de um atributo (ou objeto) que se fundamenta num domínio de aplicação. Na visão de Schreiber et al. (2002), complementada por Druziani, Kern e Catapan (2012), a tarefa de classificação é uma tarefa intensiva em conhecimento, de natureza analítica. Possui como entrada característica um objeto que, ao nível de processamento, realiza a associação entre características e classes, e como saída apresenta objetos classificados a partir de um conjunto de classes pré-definidas.

A tarefa de classificação propicia benefícios para as organizações, categorizando os seus dados internos e externos, com o intuito de complementar as informações operacionais para o apoio à decisão. Segundo Li e Tsai (2013), a classificação automática de texto é uma técnica fundamental para gerenciar grandes coleções de documentos. Ao aplicar a tarefa de classificação no contexto

organizacional, subtarefas podem ser derivadas, por exemplo, a análise de opiniões.

A Internet representa um importante meio para a divulgação das opiniões dos consumidores na forma de *posts*, *blogs*, fóruns, *sites* de empresas, entre outros. Esses relatos refletem as experiências vividas pelos consumidores em relação a serviços e produtos (PAI et al., 2013).

Percebe-se que as mídias sociais representam um importante canal de comunicação entre as organizações e os seus clientes. Segundo Chandran e Murugappan (2012), as mídias sociais são um grupo de aplicações, baseadas em Internet, que permitem a criação e a troca de informações geradas pelo usuário. Elas promovem impacto na forma de expressar opiniões e pensamentos dos consumidores e dos fornecedores de produtos e serviços. Segundo Atkinson, Salas e Figueroa (2015), a explosão no uso das mídias sociais criou uma importante oportunidade para as pessoas publicarem as suas opiniões.

Dessa forma, sendo as mídias sociais um espaço em que opiniões podem ser encontradas, as organizações precisam desenvolver ferramental capaz de capturar as visões de seus clientes, a fim de interpretá-las e auxiliar a tomada de decisão. Para isso, pode-se utilizar como base a análise de sentimentos, que é uma das possibilidades para a tarefa de classificação (BELLINI et al., 2012). A análise de sentimento tem como objetivo polarizar as informações apresentadas num texto, por exemplo, em positivas e negativas (QIU et al., 2011).

A possibilidade de se conhecer os motivos pelos quais uma sentença recebe o grau positivo ou negativo é tão importante para o apoio à decisão quanto a própria polarização (ZHANG; LIU, 2011a) (PENALVER-MARTINEZ et al., 2014).

Como exemplo, cita-se a área de *marketing*, que utiliza as informações obtidas pela análise de sentimento para saber se o lançamento de um produto, de uma nova campanha, ou mesmo, se a imagem de uma organização é boa (positiva) ou ruim (negativa), baseada na experiência do consumidor (BELLINI et al., 2012). Utiliza-se a análise de sentimento em outros domínios de aplicações a partir de informações de mídias sociais como, por exemplo, na detecção de crimes e de terrorismo (YANG; DORBIN, 2011), na análise de campanhas eleitorais (TUMASJAN et al., 2011), na área da inteligência competitiva (XU et al., 2011), entre outros.

A seção seguinte apresenta a problemática em que esta tese está centrada.

1.1 DEFINIÇÃO DO PROBLEMA

Mesmo antes da popularização da *Web*, as organizações já se preocupavam com a opinião que pessoas e entidades possuiam sobre seus produtos, serviços, atendimentos ou imagem. Segundo Feldman (2013), quando um consumidor que utiliza a Internet como canal de compra deseja adquirir um novo produto, ele tipicamente realiza buscas em fóruns e *blogs* especializados para saber opiniões sobre o produto.

Com o advento das plataformas da *Web 2.0*, o usuário da Internet deixa de ser apenas consumidor para tornar-se também produtor de conteúdo, não necessitando conhecer de técnicas de programação ou de técnicas computacionais avançadas. Isso facilita a produção e a publicação de conteúdos diversos, tais como, fotos, textos, opiniões etc. (O'RELLY, 2005, MARTINEZ; FERREIRA, 2007).

Sendo a Internet um meio democrático que permite que os usuários produzam conteúdo de maneira mais simples, observa-se uma 'explosão' no número de conteúdos desenvolvidos, o que, de certo modo, dificulta o consumo das informações por parte dos usuários que estão buscando dados nos quais possam apoiar a sua decisão (AMBINDER; MARCONDES, 2011; DURIC; SONG, 2012). Esse cenário apresenta a necessidade de recursos para classificação de textos, os quais auxiliam o gerenciamento de grandes coleções de documentos (DURIC; SONG, 2012; LI; TSAI, 2013).

Um recurso utilizado para classificação é a análise de textos, desenhos ou figuras emitidas por terceiros (LONGHI et al., 2009). Muitos pesquisadores das áreas Processamento de Linguagem Natural (PLN) e Mineração de Dados têm focado suas pesquisas na análise de sentimento ou na mineração de opinião, as quais possuem como objetivo polarizar/classificar em positivo/negativo sentenças ou documentos (QIU et al., 2011). Segundo Serrano-Gerrero et al. (2015), a análise de sentimento é um dos temas de pesquisa mais recentemente pesquisados no domínio do processamento de informação, e busca explicitar elementos referentes às opiniões, em forma de texto.

Para identificar a imagem de uma organização, pode-se utilizar a análise de sentimento, tendo como recurso, informações advindas das mídias sociais (SERRANO-GERRERO et al., 2015). A partir desses dados, a utilização de algumas técnicas se faz necessária. Segundo Liu (2010a), muitos pesquisadores têm se dedicado ao problema de classificar opiniões, como positivas ou negativas, baseados em termos, adjetivos e advérbios contidos em uma sentença.

Para alguns pesquisadores, é possível utilizar documentos relativos ao domínio de uma organização com o propósito de identificar sua imagem a partir de adjetivos e de termos relacionados, como é verifica-se no trabalho de Hatzivassiloglou e McKeowa (1997). Outros trabalhos seguem essa mesma linha, como os desenvolvidos por Wiebe (2000), Kanayama e Nasukawa (2006), Qiu et al. (2009), Bacheri, Saraee e Jong (2014).

Inicialmente, a análise, bem como a classificação de uma sentença em positiva ou negativa, é determinada pelo grau de polaridade das palavras contidas na sentença. Contudo, essa abordagem resulta em uma taxa de erro considerável, já que a classificação é sensível ao domínio e ao contexto (ZHANG; LIU, 2011a).

A análise de sentimento é vista como um recurso com potencial para avaliar informações relacionadas a uma organização por meio do monitoramento de mídias sociais (JEBASEELI; KIRUBAKARAN, 2012). Contudo, para uma análise mais efetiva, é necessário que a abordagem seja sensível ao domínio. Nesse caso, o uso de ontologias pode auxiliar (TSYTSARAU; PALPANAS, 2012; LAU; LI; LIAO, 2014; PEÑALVER-MARTINEZ et al., 2014; AGARWAL et al., 2015). A classificação de documentos apresenta desafios como, por exemplo, considerar a semântica dos conteúdos a serem classificados (CECI; WOSZEZENKI; GONÇALVES, 2014; SERRANO-GUERRERO et al., 2015).

Quando se procura polarizar uma sentença, deve-se levar em consideração a entidade de destino, ou seja, o foco do procedimento. A entidade, na polarização, é um objeto válido para o domínio em questão (LIU, 2010b; QIU et al., 2011; TSYTSARAU; PALPANAS, 2012; ATKINSON; SALAS; FIGUEROA, 2015). Para entender melhor o que é a entidade de destino e a sua importância em uma sentença, apresenta-se o seguinte exemplo: “Eu gostei deste show, a banda estava ótima, mas a acústica do estádio estava horrível, pelo menos a pista era enorme.”. Analisando a sentença de exemplo, pode-se perceber que existem quatro entidades de destino: show, banda, estádio e pista. Cada uma delas teve uma característica atribuída, e essas características se somam para identificar a polaridade da sentença.

Para cada entidade de destino, pode-se vincular uma ou mais características, que no caso, são os termos polares (termos que possuem polarização positiva ou negativa). Ainda sobre o exemplo apresentado, pode-se vincular o termo polar “gostei” com a entidade “show”, sabendo que “gostei” é um termo polar positivo. É importante observar que os termos polares não possuem classificações comuns para todos os

domínios, em muitos casos, eles podem ser ambíguos, ou seja, ter uma classificação diferente.

Para Li e Tsai (2013), a grande maioria dos algoritmos de classificação de documentos é facilmente afetada por termos ambíguos, o que pode gerar uma classificação errada. Segundo Liu et al. (2015), termos polares comumente possuem conotação (polarização) distinta, dependendo do domínio de aplicação. Essa situação pode trazer problemas para o processo de análise de sentimento.

O uso de análise de sentimento traz benefícios para a tomada de decisão de uma organização, tendo em vista que é possível mesclar a opinião de terceiros com informação internas, a fim de explicitar novos conhecimentos e/ou subsidiar a tomada de decisão (PAI et al., 2013). A utilização de classificação baseada em qualificadores, positivo e negativo, apresenta bons resultados, mas ainda está longe de proporcionar uma classificação de qualidade, ou seja, realmente representativa sobre a opinião do usuário (LI; XIA; ZHANG, 2011).

A análise de sentimento costuma focar na polarização da sentença como um todo, sem combinar o processo com as características de um produto ou dos demais objetos do domínio, o que pode tornar a classificação incompleta ou menos relevante para a tomada de decisão (WANG; XU; WAN, 2013; PEÑALVER-MARTINEZ et al., 2014; SERRANO-GUERRERO et al., 2015).

Além da orientação semântica (positiva ou negativa) não ser suficiente para uma análise mais efetiva (FENG et al., 2011), a cada nova classificação, é necessário considerar todo o processo de inferência e de classificação novamente, não aproveitando todos os raciocínios já realizados. Para que os sistemas de análise de sentimento tenham uma classificação mais efetiva, esses devem ‘aprender’ a partir das práticas que obtiveram sucesso (KAISER; SCHLICK; BODENDORF, 2011; ATKINSON; SALAS; FIGUEROA, 2015).

Wang, Xu e Wan (2013) apresentam três desafios com que toda pesquisa focada em análise de sentimento deve se preocupar:

(1) Encontrar o recurso ou o objeto nas sentenças;
(2) Obter as características relacionadas com o recurso; e
(3) Conseguir a polarização (orientação de sentimento) para a sentença

Na visão de Li e Xu (2014), pela natureza do problema da análise de sentimento, as abordagens tradicionalmente mais utilizadas pelos pesquisadores são as baseadas em estatísticas como, por exemplo: *Support Vector Machine* (SVM), *k-Nearest Neighborhood* (k-NN), entre

outras. Os autores afirmam que essas técnicas são limitadas por dois motivos:

- Frases complexas com negação ou perguntas retóricas não podem ser tratadas; e
- Informações mais detalhadas sobre a classificação, como o porquê de se receber certa definição, não podem ser obtidas.

A seção a seguir, apresenta a pergunta de pesquisa que norteia o presente trabalho.

1.2 PERGUNTA DE PESQUISA

A partir do contexto anteriormente declarado, apresenta-se a seguinte questão: Como a representação do conhecimento de determinado domínio e o armazenamento e recuperação de raciocínios passados podem auxiliar na tarefa intensiva de classificação com foco na análise de sentimento?

1.3 PRESSUPOSTOS DA PESQUISA

O presente trabalho possui alguns pressupostos que estão expostos com mais detalhes a seguir.

- Documentos não-estruturados podem abrigar informações importantes para uma organização.
- Os métodos de classificação podem possuir uma taxa de acerto melhor se forem sensíveis ao domínio de aplicação.
- É possível utilizar dados e informações presentes em recursos da *Web 2.0* para auxiliar na complementação das bases de conhecimento de uma organização.
- A análise de sentimento pode representar a imagem de um produto ou serviço baseando-se em textos e em sentenças extraídos da Internet.
- É possível armazenar a inferência de uma classificação passada para utilizá-la em uma nova classificação.
- A identificação de características de um produto ou de um objeto em uma sentença pode auxiliar o entendimento do resultado de uma classificação.
- O uso de ontologias de domínio promove uma contextualização aprimorada de recursos (textos) e, portanto, uma melhor polarização/classificação.
- Uma inferência pode ser representada na forma de árvore, de modo que seja possível armazená-la, recuperá-la e reaproveitá-la na tarefa de classificação.

1.4 OBJETIVOS DO TRABALHO

Nesta seção, encontram-se descritos os objetivos geral e específicos que se busca atingir ao longo do desenvolvimento desta tese.

1.4.1 Objetivo geral

Desenvolver um modelo em que seja possível representar o conhecimento de domínio, bem como, armazenar e recuperar raciocínios passados para suportar a tarefa de classificação voltada à polarização de conteúdo não-estruturado.

1.4.2 Objetivos específicos

Para este trabalho, formularam-se os seguintes objetivos específicos:

- Identificar e implementar técnicas de Engenharia do Conhecimento para a aplicação de classificadores semânticos no processo de análise de sentimento;
- Analisar o uso de uma abordagem baseada em casos passados para auxiliar na solução proposta;
- Propor um modelo que utilize os elementos explicitados e aborde as oportunidades de pesquisa identificadas durante a revisão sistemática;
- Demonstrar a viabilidade do modelo proposto, por meio da construção de um protótipo, assim como a aplicação deste em alguns cenários;
- Realizar uma análise comparativa, focada na análise de sentimento, com outros modelos de classificação de texto;
- Avaliar a robustez do modelo proposto comparando-o com classificadores tradicionais por meio de medidas padronizadas.

Na próxima seção, apresenta-se a justificativa e a relevância do tema.

1.5 JUSTIFICATIVA E RELEVÂNCIA DO TEMA

As organizações estão, cada vez mais, transformando-se em instituições do conhecimento, pois suas atividades são de natureza cognitiva, ou seja, são atividades intensivas em conhecimento (LYTRAS; POULOUDI, 2006). Essas organizações têm se preocupado com a criação de conhecimento organizacional que, segundo Nonaka e Krogh (2009), é o processo de tornar o conhecimento criado por

indivíduos disponível, de tal forma que seja possível gerenciá-lo e utilizá-lo como suporte para a construção de sistemas baseados em conhecimento.

Os documentos de uma organização podem trazer muito ganho ao processo de apoio à decisão, já que abrigam informações do seu domínio de aplicação. Esses registros podem também ser utilizados no processo de construção de bases de conhecimento, como apresentado nos trabalhos de El Sayed e Hacid (2008), Fortuna, Lavrac e Velardi (2008), Ceci et al. (2010) e Ceci, Pietrobon e Gonçalves (2012).

Além do uso dos documentos da instituição, podem-se utilizar informações externas à organização para agregar valor ao processo de explicitação e de criação do conhecimento organizacional, bem como é possível utilizá-las diretamente para apoiar a tomada de decisão organizacional. Para Liu (2010b), no passado, quando uma pessoa precisava tomar alguma decisão, tipicamente perguntava a opinião de seus amigos e familiares. Uma organização, quando queria encontrar uma opinião sobre os seus produtos e serviços, recorria aos seus clientes. Antes da popularização da Internet, as organizações tinham de apelar a canais como o telefone e as malas diretas contendo formulários de sugestões e reclamações. Nos dias atuais, as mídias sociais, bem como as demais aplicações da *Web 2.0*, podem ser canais para encontrar essas opiniões (DURIC; SONG, 2012).

As mídias sociais apresentam informações em texto livre, ou seja, de forma não-estruturada (ATKINSON; SALAS; FIGUEROA, 2015). Dessa forma, o uso da análise de sentimento torna-se uma alternativa na extração de informações sobre produtos e serviços a fim de adicionar valor aos dados internos da organização (JEBASEELI; KIRUBAKARAN, 2012; KONTOPOULOS et al., 2013).

Segundo Liu (2010b), com o passar dos anos, tanto a academia quanto as organizações têm voltado os olhos para a análise de sentimento, uma vez que as visões exteriores à instituição são estratégicas à tomada de decisão. Para He, Alani e Zhou (2010), a análise de sentimentos visa compreender a emoção subjetiva da informação, tais como opiniões, atitudes e anseios expressos no texto.

Tornou-se um tema frequentemente estudado nos últimos anos, principalmente por causa da explosão no uso de dispositivos móveis e do aumento do tempo de exposição dos usuários à Internet, o que permite que sejam facilmente publicadas opiniões nos meios de comunicação social, incluindo *blogs*, fóruns de discussão, *tweets* etc. (DURIC; SONG, 2012; MEDHAT; HASSAN; KORASHY, 2014). As organizações devem considerar o uso das informações obtidas por meio

de recursos da *Web 2.0* (ATKINSON; SALAS; FIGUEROA, 2015). Além disso, é necessário que esses dados (opiniões) sejam cruzados com elementos da base de conhecimento para que se tenha uma classificação mais eficiente e coerente com o domínio de aplicação (KONTOPOULOS et al., 2013).

O domínio de aplicação deve ser modelado de modo que seja possível armazená-lo e utilizá-lo em sistemas de conhecimento, reaproveitando-o sempre que necessário (PEÑALVER-MARTINEZ et al., 2014). Para Nassirtoussi (2014), o uso de ontologias como suporte à análise de sentimento é uma abordagem interessante, pois apresenta elementos do domínio da instituição durante o processo de classificação.

Segundo Tsytsarau e Palpanas (2012), as ontologias podem trazer benefícios à análise de sentimento, tanto na identificação de recursos e características de um produto ou serviço presente na sentença quanto na própria polarização, a partir da identificação de termos polarizados do domínio.

A identificação da entidade de destino é um dos principais desafios do processo de análise de sentimento (ZHAO et al., 2015). Vários trabalhos utilizam as ontologias para auxiliar na identificação de conceitos e entidades relacionados com as sentenças, como, por exemplo, Penãlver-Martinez et al. (2014), Lau, Li e Liao (2014), Liu et al. (2015), entre outros.

A partir da identificação do domínio de aplicação, é possível tratar os termos polares ambíguos, ou seja, que possuem polarização distinta dependendo da sua aplicação (LIU et al., 2015, AGARWAL et al., 2015). Conforme afirmam Li e Tsai (2013), a capacidade para tratar a ambiguidade de um classificador está diretamente ligada à tarefa de se classificar com precisão, o que demonstra a importância de levar em consideração informações do domínio de aplicação.

Com o uso de classificadores semânticos combinados aos elementos tradicionais da análise de sentimento, organizações do conhecimento têm uma poderosa fonte de registros para valorizar a sua base de conhecimento, podendo auxiliar ainda mais na tomada da decisão e na descoberta de conhecimento (LI; TSAI, 2013; MEDHAT; HASSAN; KORASHY, 2014).

O uso de métodos baseados em casos ampara o armazenamento, a aprendizagem e a recuperação de episódios passados para que possam auxiliar na classificação de novos casos. Essa abordagem é utilizada no contexto da análise de sentimento em alguns trabalhos, entre eles, Zhang e Liu (2011a), Ohana et al. (2012), Minhas et al. (2013), Sani et al. (2013), Dong et al. (2013a) e Dong et al. (2013b).

Por fim, em muitos casos, o caminho percorrido para chegar até a orientação semântica é mais importante que a própria classificação, ou seja, saber o porquê de se ter atingido determinada informação e quais outros qualificadores estão envolvidos na análise é essencial para que, posteriormente, se possa utilizar esse raciocínio (inferência) em novas classificações. Trabalhos como os propostos por Kaiser, Schlick e Bodendorf (2011) e Li e Xu (2014) corroboram tais afirmações.

1.6 ORIGINALIDADE

O modelo proposto neste trabalho procura combinar a técnica de raciocínio baseado em casos (RBC) com o uso de ontologia para auxiliar no processo de classificação semântica, mais precisamente na análise de sentimento, de modo que todo o raciocínio já desenvolvido seja armazenado e reaproveitado em novas classificações.

Utilizou-se, como ponto de partida, uma revisão sistemática para identificar como as áreas da classificação semântica e da análise de sentimento estão organizadas e quais são suas linhas de pesquisa. Durante o processo da revisão sistemática, não foram encontrados artigos que tivessem como foco o processo de armazenamento, recuperação e reutilização de raciocínios passados para o processo de análise de sentimento, o que demonstra uma lacuna enquanto proposta de solução para a tarefa de classificação.

Dos 85 artigos classificados como relevantes e lidos, apenas 7 tratam do tema-raciocínio relativo à análise de sentimento: Wu et al. (2010), Huang e Qiu (2010), Cambria et al. (2012a), Cambria et al. (2012b), Li e Tsai (2013), Pai (2013), Kontopoulos et al. (2013).

As escolhas dos elementos do modelo proposto estão diretamente ligadas aos já apresentados na revisão sistemática. Neste modelo se combinaram técnicas de classificação com RBC e com ontologia. Essa forma de orquestrar as soluções não foi encontrada em nenhum artigo presente na revisão sistemática.

A ideia de utilizar RBC partiu da análise do trabalho de Li e Tsai (2013), que combina essa técnica com a Lógica *Fuzzy* para realizar a análise de sentimento. O ponto em que o presente modelo diverge dessa abordagem é que os casos armazenados são os próprios documentos já polarizados e não o raciocínio que se teve para chegar à classificação.

Sobre o uso da ontologia, muitos artigos a utilizam como forma de armazenar o conhecimento de domínio da organização, bem como para representar o domínio da análise de sentimento. No contexto de uso das ontologias para raciocínio, quatro artigos apresentam explicitamente o seu uso: CAMBRIA et al., 2012a; 2012b; PAI, 2013; e

KONTOPOULOS et al., 2013. Nesses artigos, a ontologia foi utilizada como forma de representação de conhecimento e como subsídio para os raciocínios. Contudo, em nenhum dos trabalhos, o armazenamento ou o reaproveitamento de raciocínios anteriores foi utilizado.

Após a conclusão da etapa de experimentos deste trabalho, optou-se por revisitar as bases indexadas de artigos e atualizar a revisão sistemática. Constatou-se que, até o momento, o *gap* identificado a partir da literatura ainda continua aberto. Mais informações são apresentadas na Seção 4.5.1.

1.6.1 Contribuições

A contribuição principal deste trabalho consiste na concepção de um novo modelo de classificação semântica focado no reaproveitamento dos raciocínios (inferências) efetuados a partir da utilização de RBC com ontologias para a classificação de sentenças ou documentos novos.

Além da contribuição principal, a seguir são apresentadas outras possíveis contribuições do referente trabalho.

- Definição de uma estrutura de dados para armazenamento, recuperação e reaproveitamento de inferências efetuadas durante o processo de classificação.
- Extensão de uma ontologia de domínio para a utilização no contexto da análise de sentimento.
- Implementação de um protótipo que ateste a viabilidade do modelo proposto, bem como a sua aplicação em estudos de caso.
- Elaboração de uma metodologia que apresente os conceitos do domínio e uma justificativa para o sentimento, a qual pode ser utilizada também como validação para o processo de classificação.
- Definição de um método para adequação do léxico de sentimento pelos casos passados.
- Implementação de uma estratégia para tratamento de ambiguidade.
- Método para construção de uma base de conhecimento para uma organização baseado nas classificações de sentenças.

A próxima seção demonstra o escopo da presente pesquisa.

1.7 ESCOPO DO TRABALHO

Este trabalho tem como objetivo a classificação de textos, identificando sentimentos e opiniões presentes em bases textuais

disponíveis na *web*, não tendo como foco a extração e a análise de sentimentos a partir de texto disponíveis em revistas ou outras mídias, sejam elas, impressas, sonoras ou imagéticas.

O modelo proposto, utiliza-se de uma série de técnicas e métodos, como o Raciocínio Baseado em Casos, o processamento de linguagem natural e os métodos estatísticos, não tendo como pretensão trazer uma contribuição direta para essas ferramentas.

Uma ontologia de domínio foi utilizada como base de conhecimento para o modelo. Contudo, não é o foco desta pesquisa apresentar técnicas para a modelagem ou criação da ontologia em questão, tendo em vista que ela pode ser substituída por uma que faça mais sentido ao domínio em questão.

Também não se objetiva que o protótipo desenvolvido a partir do modelo proposto apresente características de um produto final, uma vez que o protótipo tem como foco a avaliação do modelo.

Para a avaliação do modelo proposto, serão utilizados apenas textos curtos, de no máximo duas páginas, não sendo o foco identificar a polarização de livros ou textos longos. Este trabalho também não tem como objetivo contribuir para as áreas de tratamento de negação e análise de subjetividade.

A próxima seção tem como objetivo apresentar a metodologia de pesquisa utilizada neste trabalho.

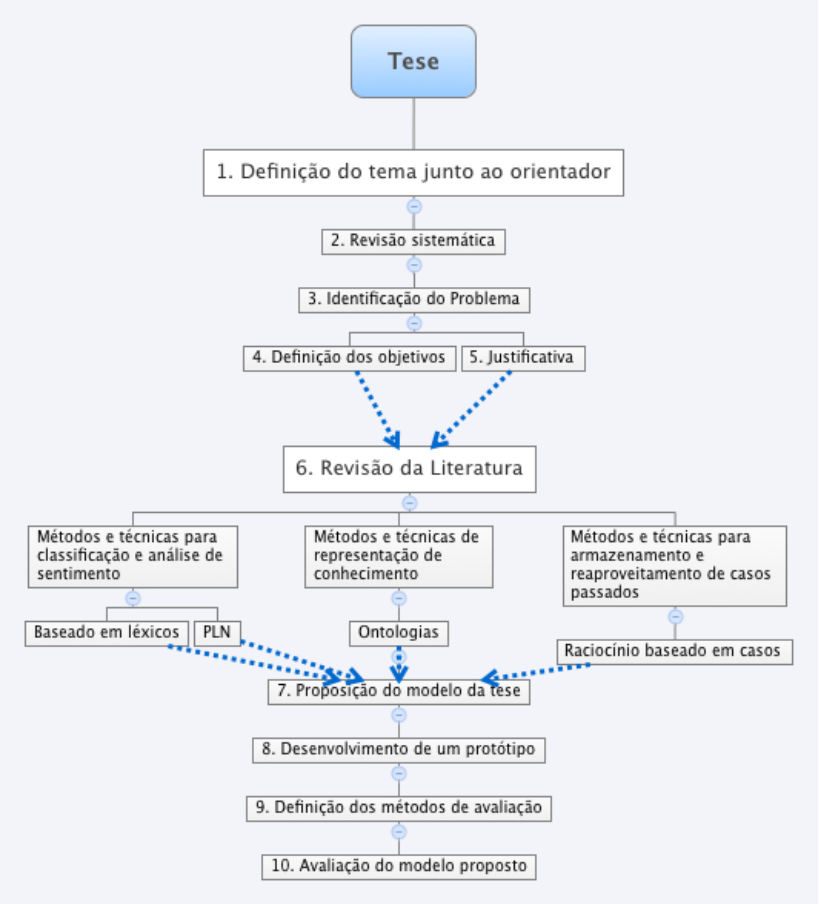
1.8 METODOLOGIA DA PESQUISA

A presente seção visa descrever a metodologia utilizada neste trabalho a fim de classificar a pesquisa dentro dos diversos pontos de vista. Segundo Gil (1999, p.42, apud SILVA; MENEZES, 2001, p.19), “o objetivo fundamental da pesquisa é descobrir respostas para problemas mediante o emprego de procedimentos científicos.”.

O trabalho aqui apresentado, sob o ponto de vista de sua natureza, é caracterizado como uma pesquisa aplicada, a qual, conforme Silva e Menezes (2001, p.20), “objetiva gerar conhecimentos para aplicação prática, dirigidos à solução de problemas específicos. Envolve verdades e interesses locais.”.

Para atingir os objetivos desta pesquisa, o trabalho dividiu-se nas etapas apresentadas na Figura 1:

Figura 1 – Etapas metodológicas do trabalho



Fonte: Elaborado pelo autor

Inicialmente, definiu-se o tema do trabalho com o orientador (1). Após a definição, formulou-se um protocolo para aplicar durante o processo da revisão sistemática (2) – mais detalhes sobre essa etapa podem ser observados na Seção 2.1.1.

A etapa de revisão sistemática é importante para demonstrar lacunas existentes na área de classificação de textos e de análise de sentimento. A partir dessa etapa, definiu-se o problema de pesquisa (3), os objetivos (4) e a justificativa (5) para a presente pesquisa.

Após as etapas anteriores serem finalizadas, desenvolveu-se a revisão da literatura (6). Primeiramente, definiu-se conceitos

importantes para a tese, por exemplo: sentimento, opinião, classificação semântica e análise de sentimento. O próximo passo foi identificar os principais métodos e técnicas aplicados para a tarefa de classificação, que foram explicitados durante a revisão sistemática. Na sequência, foram definidos conceitos sobre ontologia como forma de armazenamento e de reaproveitamento de conhecimento do domínio. Por fim, foi identificado um método ou técnica para armazenar e reaproveitar casos passados.

Com a revisão da literatura concluída, definiu-se uma primeira proposta de modelo para a tese (7). A partir do modelo definido, foi desenvolvido um protótipo (8), a fim de atestar a viabilidade do modelo.

Na etapa (9), procurou-se definir o método e as medidas de avaliação do modelo proposto a partir da execução de um experimento sobre o protótipo desenvolvido.

Por fim, avaliou-se os resultados (10), de forma que fosse possível compará-los com outros modelos e com propostas similares para solucionar o problema de pesquisa.

1.9 ADERÊNCIA AO OBJETO DE PESQUISA DO PROGRAMA

Esta seção tem como objetivo demonstrar a aderência da tese ao objeto de pesquisa do Programa. Para isso, as três subseções seguintes apresentam a identidade da tese, o contexto estrutural no EGC e as referências factuais.

1.9.1 Identidade

A aderência deste trabalho ao objeto de pesquisa do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento[1] pode ser reforçada a partir do objeto de pesquisa e dos objetivos do Programa:

> O objeto de pesquisa do EGC refere-se aos macroprocessos de explicitação, gestão e disseminação do conhecimento. Estes incluem os processos de criação (e.g., inovação de ruptura), descoberta (e.g., redes sociais), aquisição (e.g., inovação evolutiva), formalização/codificação (e.g., ontologias), armazenamento (e.g., memória organizacional), uso (e.g., melhores práticas), compartilhamento (e.g., comunidades de prática), transferência (e.g., educação corporativa) e evolução (e.g., observatório do conhecimento). [...] Deste modo, o objetivo do EGC consiste em

[1] Disponível em: <http://www.egc.ufsc.br/htms/vermais_index.htm>. Acesso em: 10 out. 2010.

> investigar, conceber, desenvolver e aplicar modelos, métodos e técnicas relacionados tanto a processos/bens/serviços como ao seu conteúdo técnico-científico [...].

Pode-se observar que o modelo está aplicado sobre o processo de aquisição do conhecimento disponível nas mídias sociais, a fim de possibilitar a convergência das informações internas com as externas à organização.

O conhecimento, no contexto deste trabalho, está presente na representação do domínio de aplicação, pelo uso das ontologias, e na formalização e armazenamento dos raciocínios (inferências) passados na forma de árvores de sentimento.

1.9.2 Contexto estrutural no EGC

A Engenharia do Conhecimento nasceu de um ramo da Inteligência Artificial (IA), tendo como objetivo estudar técnicas e métodos para a extração, manipulação e classificação do conhecimento, promovendo suporte à construção de sistemas de conhecimento, bem como fornecendo insumos para a Gestão do Conhecimento (SCHREIBER et al., 2002; STUDER; BENJAMINS; FENSEL, 1998).

No que tange ao escopo deste trabalho, o aspecto que o contextualiza na área de Engenharia do Conhecimento reside no fato de o modelo ter como objetivo a materialização, principalmente dos macroprocessos de explicitação, do conhecimento, sem perder a possibilidade de promover suporte aos macroprocessos de gestão e disseminação do conhecimento.

Segundo Levy (2009), o ferramental disponibilizado pela *Web 2.0* é um importante aliado para a gestão do conhecimento, pelo fato de os seus usuários trocarem conhecimento diretamente. Diante das informações disponibilizadas em ferramentas colaborativas da *Web 2.0*, as organizações podem analisar opiniões sobre seus produtos, serviços ou imagem.

Para que os dados extraídos possam somar valor às bases de caso e conhecimento, é necessário anteriormente submetê-las a um processamento em que elas sejam qualificadas. Para tal, é utilizado ferramental de processamento de linguagem natural, análise de sentimento, elementos de análise da linguística e da inteligência artificial.

O presente trabalho está centrado na linha de pesquisa Teoria e Prática em Engenharia do Conhecimento, que busca, por ferramental computacional, apoiar a processos de aquisição e de representação do

conhecimento, permitindo que esses processos possam ser geridos e utilizados pelas três áreas do programa: engenharia, gestão e mídia do conhecimento.

1.9.3 Referências factuais

O presente trabalho tem como foco principal a área da classificação semântica, mais precisamente a área de análise de sentimento. A seguir, são apresentados os dois trabalhos desenvolvidos no programa que são relacionados com a área de classificação semântica.

PIZZOL, Leandro Dal. Uso da Web de Dados como Fonte de Informação no Processo de Inteligência Competitiva Setorial. Dissertação, 2014.

RIBEIRO, Samuel F. Sistema de Conhecimento para Gestão Documental no Setor Judiciário: uma aplicação no Tribunal Regional Eleitoral de Santa Catarina. Dissertação, 2010.

Existe apenas uma tese que comenta sobre a temática de sentimento no seu contexto, qual seja:

GARCIA, Roseli Amado S. Mídias do Conhecimento na autoconstrução de sujeitos complexos: um estudo de caso no Museu de Arte Moderna da Bahia. Tese, 2010.

Sobre o tema Análise de Sentimento ou Mineração de Opinião, existe apenas uma dissertação, apresentada a seguir.

LINDNER, Luis Henrique. Diretrizes para o design de interação em redes sociais temáticas com base na visualização do conhecimento. Dissertação, 2015.

Para obter os textos aplicados na análise de sentimento, de modo que seja possível coletar as opiniões de usuários ou clientes sobre seus produtos ou serviços, é necessário utilizar como base plataformas disponíveis na *Web 2.0*. Sobre *Web 2.0*, existem seis trabalhos já defendidos no Programa, como é possível observar a seguir.

DRUZIANI, Cássio Frederico Moreira. O Repositório Web Como Potencializador Do Conhecimento Em Objetos De Aprendizagem. Tese, 2014

DZIEKANIAK, Gisele Vasconcelos. Método para Inclusão de Conhecimento Presente em Mídias Sociais no Aprimoramento de Plataformas de Governo Eletrônico. Tese, 2012.

BENÍTEZ HURTADO, Segundo Raymundo. Práticas de Gestão do Conhecimento no Processo de Formação de Docentes em uma Universidade Privada do Equador. Dissertação, 2012.

BEIRÃO FILHO, José Alfredo. Criação e Compartilhamento do Conhecimento da Área de Moda em Um Sistema Virtual Integrado – SIMODA. Tese, 2011.

OLIVEIRA, Thiago P. S. de. Sistemas Baseados em Conhecimento e Ferramentas Colaborativas para a Gestão Pública: Uma Proposta ao Planejamento Público Local. Dissertação, 2009.

SILVA, Rodrigo Gecelka da. O Potencial Educacional dos Mundos Virtuais Tridimensionais: Um Estudo de Caso do Second Life. Dissertação, 2012.

Além da busca por trabalhos relacionados com as temáticas da tese, objetiva-se identificar trabalhos que utilizaram como base para encontrar o problema de pesquisa, revisões sistemáticas. Foram encontrados vinte e um trabalhos já defendidos, como é possível observar a seguir.

HELOU, Angela Regina Heinzen Amin. Avaliação da Maturidade da Gestão do Conhecimento na Dministração Pública. Tese, 2015.

MEZZAROBA, Mariana Pessini. Requisitos para a Avaliação de Portais de Governo Eletrônico do Poder Judiciário a partir das Resoluções de Metas do CNJ. Dissertação, 2015.

REGINALDO, Thiago. Referenciais Teóricos e Metodológicos para a Prática do Design Thinking na Educação Básica. Dissertação, 2015.

SÁ, Marcelo Alexandre. Redes De Cooperação Como Estratégia Para Desenvolvimento Da Agricultura Familiar: Programa SC Rural. Dissertação, 2014.

LINO, Sônia Regina Lamego. Diretrizes para a Institucionalização da Gestão do Conhecimento na Rede Federal de Educação Profissional, Científica e Tecnológica, Brasil. Tese, 2013.

BERG, Carlos Henrique. Avaliação de Ambientes Virtuais de Ensino Aprendizagem Acessíveis Através de Testes de Usabilidade com Emoções. Dissertação, 2013.

MUÑOZ, Denise Leonora Cabrera. Processos de Conhecimento Associados à Gestão para Sustentabilidade: Um Estudo Baseado na Revisão Sistemática de Literatura. Dissertação, 2013.

BONILLA, Maria Alejandra Maldonado. Recompensas e Retenções de Profissionais Voltados para Atividades de Conhecimento em Organizações. Dissertação, 2013.

SANTANA, Julival Queiroz de. Liderança Autêntica no Batalhão de Operações Policiais Especiais de Santa Catarina. Dissertação, 2012.

DIAS, Adriano Júnior. Relações entre a Estrutura Organizacional, a Gestão do Conhecimento e a Inovação, em Empresas de Base Tecnológica. Dissertação, 2012.

BRITO, Ronnie Fagundes de. Modelo de Referência para Desenvolvimento de Artefatos de Apoio ao Acesso dos Surdos ao Audiovisual. Tese, 2012.

MALDONADO, Mauricio Uriona. Dinâmica de Sistemas Setoriais de Inovação: Um Modelo de Simulação Aplicado no Setor Brasileiro de Software. Tese, 2012.

SCHNEIDER, Elton Ivan. Uma Contribuição aos Ambientes Virtuais de Aprendizagem (AVA) Suportados pela Teoria da Cognição Situada (TCS) para Pessoas com Deficiência Auditiva. Dissertação, 2012.

SCHMITZ, Ana Lúcia Ferraresi. Competências Empreendedoras: Os Desafios dos Gestores de Instituições de Ensino Superior como Agentes de Mudança. Tese, 2012.

SEWALD JUNIOR, Egon. Modelagem de Sistema de Conhecimento para Apoio a Decisão Sentencial na Justiça Estadual. Dissertação, 2012.

ROCHA, Paula Regina Zarelli. Métodos de avaliação de ativos intangíveis e capital intelectual: análise das competências individuais. Dissertação, 2012.

SILVEIRA, Rosana Rosa. Diretrizes para mitigar as barreiras à implementação da gestão do conhecimento em organizações. Tese, 2011.

CABRAL, Rodrigo Bittencourt. Concepção, implementação e validação de um enfoque para integração e recuperação de conhecimento distribuído em bases de dados heterogêneas. Dissertação, 2010.

GIGLIO, Kamil. Análise comparativa entre IPTV, WEBTV e TVD com foco em disseminação do conhecimento. Dissertação, 2010.

MACHADO, Cátia dos Reis. Análise estratégica baseada em processos de Inteligência Competitiva (IC) e Gestão do Conhecimento (GC): proposta de um modelo. Tese, 2010.

KESSLER, Nery Ernesto. Revisão sistemática e metanálise da acurácia diagnóstica de testes laboratoriais para giardíase: contribuição para a gestão do conhecimento. Dissertação, 2007.

Diante das referências apresentadas, percebe-se que o presente trabalho está de acordo com a área de concentração da Engenharia do Conhecimento e possui trabalhos prévios que abordam temáticas similares.

1.10 ESTRUTURA DO TRABALHO

Este trabalho é composto por cinco capítulos, além da introdução que aqui se apresenta, sendo os demais relacionados a seguir.

- O Capítulo 2 é composto por um referencial teórico no qual se apresentam as áreas: classificação, análise de sentimento, ontologias e raciocínio baseado em casos.
- No Capítulo 3, apresenta-se o modelo proposto.
- No Capítulo 4, é apresentada toda a evolução do modelo de tese até a sua versão final, incluindo a proposição de avaliação do modelo por meio da discussão dos resultados alcançados pelos experimentos e também por uma análise

comparativa com modelos correlatos ao proposto neste trabalho.

- O Capítulo 5 e último, apresenta as considerações finais do trabalho, os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem como objetivo apresentar o referencial teórico dos temas que são imprescindíveis para o desenvolvimento deste trabalho. Inicialmente, é apresentado o estado da arte, proveniente de uma revisão sistemática que utiliza como base termos relacionados ao tema da tese.

A seção seguinte apresenta um referencial breve sobre a classificação de texto. Em seguida, é estudado o tema análise de sentimento, seção em que são apresentadas as suas subáreas bem como as principais técnicas e abordagens disponíveis.

O foco da tese está centrado na tarefa intensiva de conhecimento da classificação. Segundo Schreiber et al. (2002), classificação consiste na vinculação de objetos a classes, levando em consideração as suas características. Os autores ainda afirmam que essa é uma etapa analítica.

A classificação pode ser aplicada a diferentes domínios. No presente trabalho, pretende-se classificar sentenças e documentos como positivo ou negativo, limitando o uso da classificação à área da análise de sentimentos. Para ser mais eficiente, a classificação deve levar em consideração o domínio do problema em que está sendo aplicada, para isso o uso de ontologias pode trazer uma série de benefícios.

Depois da seção sobre análise de sentimento, é apresentada uma revisão sobre ontologia e, por fim, sobre raciocínio baseado em casos.

2.1 ESTADO DA ARTE

Para a construção desta seção, utilizou-se a revisão sistemática. Segundo Melgar Sasieta (2011), as revisões sistemáticas não são iguais às revisões narrativas. No caso da primeira, um protocolo deve ser definido e seguido. A revisão sistemática está organizada em três etapas, uma de planejamento, que apresenta como pretende-se organizar e executar a revisão. A segunda etapa representa a execução da revisão. Nessa etapa, o objetivo é descrever como foi o processo e de que modo chegou-se ao resultado. Na terceira etapa, são ilustrados os resultados alcançados.

2.1.1 Etapa 1: Planejamento da revisão

Planejou-se fazer três buscas distintas na fonte de dados, com objetivos diferentes, a fim de dar suporte à construção do referencial bibliográfico e para auxílio das informações presentes no Capítulo 1.

- **Busca 1**: levantamento da área de classificação semântica e análise de sentimento.

    - Objetiva-se, com essa busca, identificar como a área está organizada, quais são os seus desafios e as suas propostas de soluções.
- **Busca 2**: palavras-chave da problemática.
    - Essa busca tem como foco encontrar as possíveis soluções para as áreas relacionadas à problemática descrita, a fim de encontrar o *gap* nas propostas de soluções.
- **Busca 3**: palavras-chave da proposta de solução.
    - Objetiva-se, com essa busca, verificar se a proposta de solução pode ser considerada inédita e original.

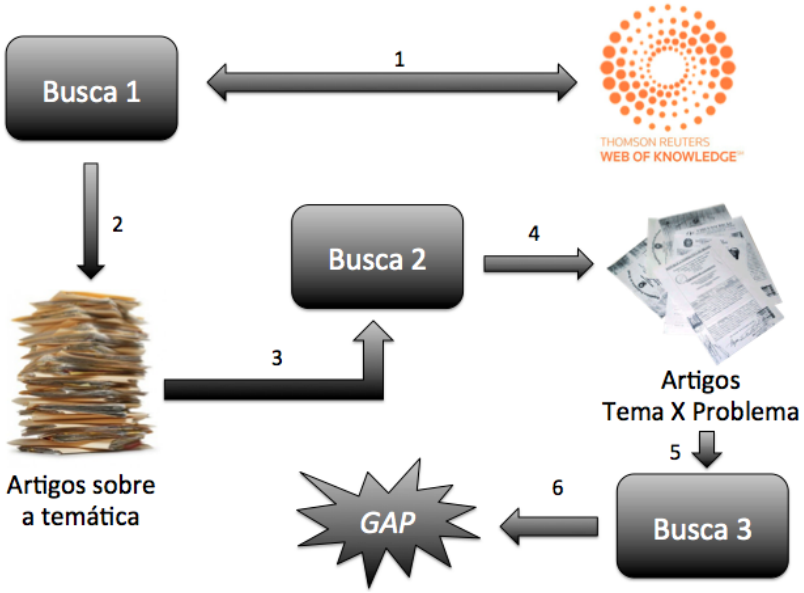
A temática definida para a revisão sistemática está focada em ‘classificação semântica’ e ‘análise de sentimento’. Como objetivo tem-se a obtenção de artigos científicos relacionados ao tema e disponíveis em bases de conhecimento.

Após definir a temática e o objetivo da revisão, formulou-se o seguinte questionamento como pergunta da revisão: Como armazenar e utilizar raciocínios de polarização passados para auxiliar na classificação semântica de novas sentenças?

O próximo passo foi a definição da fonte de dados, que, por conta do convênio que a UFSC/CAPES possui, optou-se pela base *Web of Knowledge | Web of Science*, tendo também em vista a grande quantidade de artigos disponíveis e a abrangência de áreas suportadas.

Para a execução da busca, definiu-se a estratégia apresentada com mais detalhes na Figura 2.

Figura 2 – Passos para a execução das buscas



Fonte: Elaborado pelo autor

A Figura 2 apresenta todos os passos necessários para a execução dos três tipos de busca definidos. Cada seta da figura representa um passo, conforme descrição a seguir.

- **Passo 1**: a partir dos termos de busca, são selecionados todos os artigos disponíveis na base escolhida. Nesse caso, a *Web of Knowledge*.
- **Passo 2**: os artigos são separados em relevantes e não relevantes para a temática. Depois, todos os artigos classificados como pertinentes são lidos e têm extraídas as suas características principais.
- **Passo 3**: esse passo representa a busca por artigos que façam parte da temática, mas que também estejam relacionados com o problema de pesquisa. Por conta disso, não são efetuadas mais consultas à base de artigos e sim, aos artigos já recuperados.
- **Passo 4**: os artigos retornados do processo de busca 2 são lidos e classificados pela sua relação com o tema e o problema. A partir desse ponto, é verificado se existem trabalhos que já tratam do problema, de forma a saber se ele é inédito.

- **Passo 5**: é formulada uma proposta de solução e são selecionados os artigos recuperados pelo processo de busca 2.
- **Passo 6**: para encontrar o *gap*, é verificado se existe algum trabalho que se propõe a resolver o problema de pesquisa, utilizando a mesma hipótese desta tese.

A próxima seção apresenta o processo de execução das três buscas propostas, documentando os elementos de cada etapa.

2.1.2 Etapa 2: Execução da revisão

Esta seção relata todo o processo de busca efetuado durante a etapa de revisão sistemática. Foram definidos três tipos de buscas distintas, cada qual é apresentada em seção própria. A seguir, são apresentados os elementos da primeira busca, na qual o alvo é a base *Web of Knowledge*.

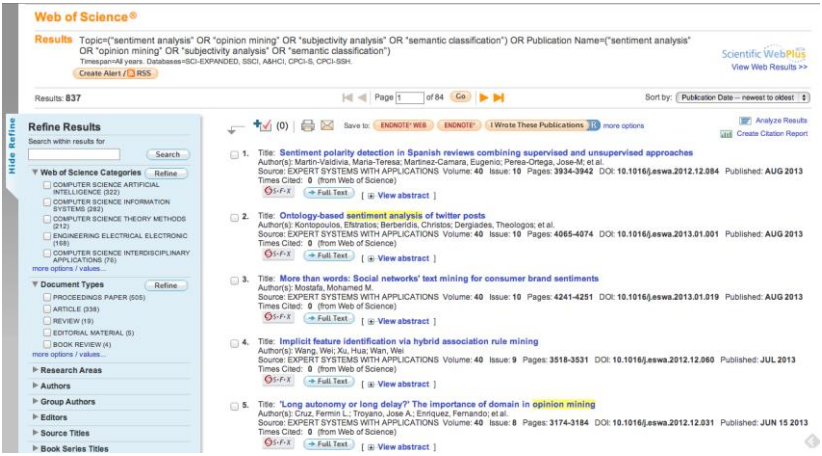
2.1.2.1 Busca 1 – *Web of Knowledge*

Primeiramente, especificou-se os termos de busca relacionados ao tema da pesquisa. Os termos escolhidos foram: análise de sentimento; mineração de opinião; análise de subjetividade; e classificação semântica. O conector entre os termos utilizados para a busca foi "OU". Com isso, pretendeu-se recuperar todos os artigos relacionados aos quatro termos de buscas selecionados. Vale lembrar que a base escolhida tem, na sua grande maioria, artigos em inglês.

A consulta montada ficou da seguinte maneira: *"semantic analysis" OR "opinion mining" OR "subjectivity analysis" OR "semantic classification".*

Os campos de busca selecionados foram *"Topic" OR "Publication name"*. A Figura 3, a seguir, apresenta a tela em que a busca foi efetuada.

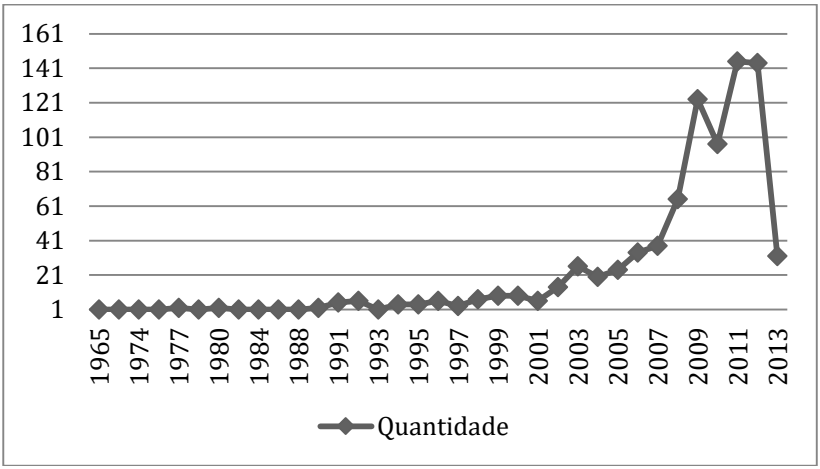
Figura 3 – Portal *Web of Science*



Fonte: Elaborado pelo autor

A busca retornou 837 artigos. Formulou-se a Figura 4 para demonstrar a distribuição dos artigos por ano.

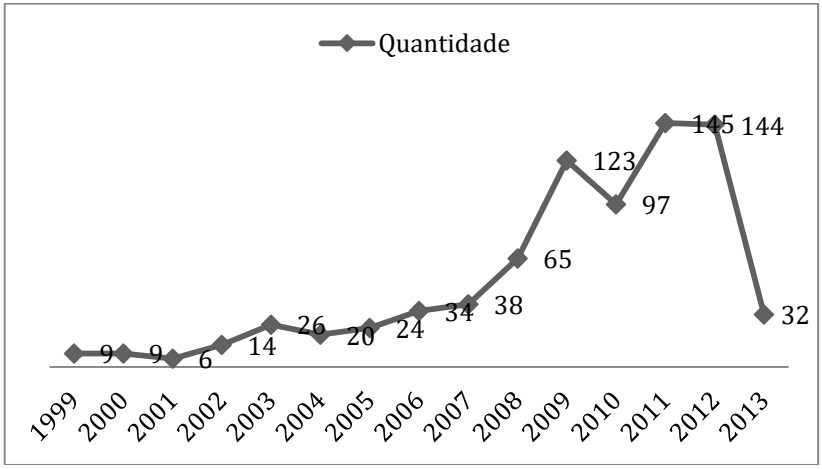
Figura 4 – Artigos retornados da busca distribuídos por ano



Fonte: Elaborado pelo autor

Como é possível observar na Figura 4, a partir de 1999, houve um aumento na publicação de artigos relacionados ao tema do presente trabalho. A Figura 5 apresenta os artigos publicados entre 1999 e meados de 2013.

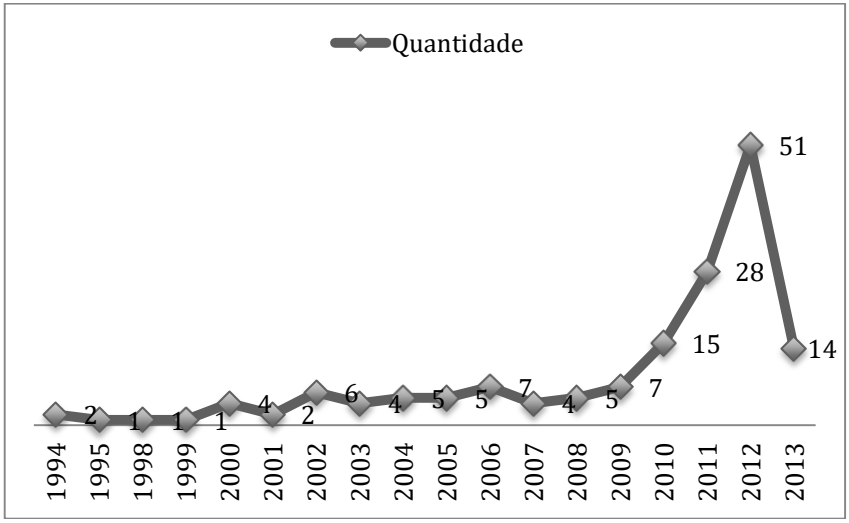
Figura 5 – Artigos publicados entre 1999 e meados de 2013



Fonte: Elaborado pelo autor

Dos 837 artigos resultantes da pesquisa efetuada na *Web of Knowledge*, apenas 166 estavam disponíveis para *download*, utilizando como alicerce o convênio da UFSC-CAPES com a base em questão. A Figura 6 expõe a distribuição dos artigos coletados por ano.

Figura 6 – Artigos coletados por ano



Fonte: Elaborado pelo autor

É importante ressaltar que existe uma queda na publicação de artigos sobre o tema em 2013, mas o motivo para isso é que esta revisão sistemática foi executada no meio do ano de 2013.

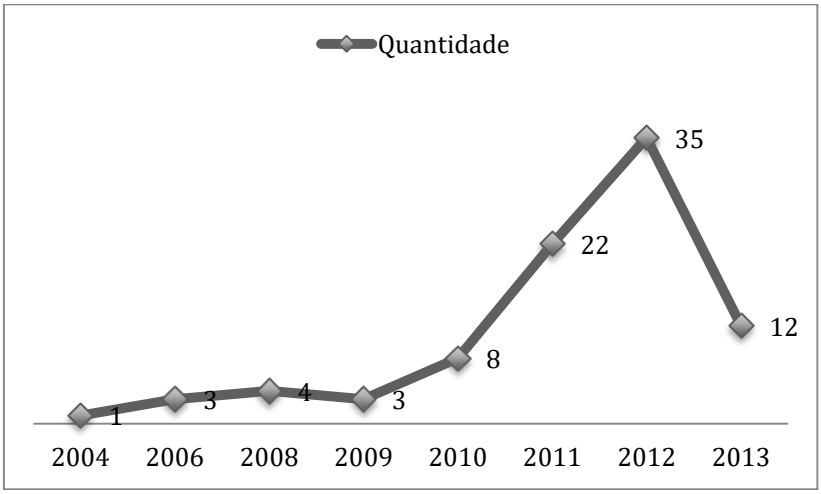
Para a seleção dos artigos recuperados, utilizou-se como base para a avaliação, o título do artigo, seu resumo e sua introdução. Considerou-se interessante para o presente trabalho apenas os artigos que apresentam análises e proposições focadas nas áreas da Computação e da Engenharia do Conhecimento, não sendo considerados os trabalhos que procuram verificar a análise de sentimento ou a mineração de opinião na perspectiva da Saúde ou das Ciências Sociais.

A partir da leitura do resumo e da introdução, foram selecionados 85 artigos como relevantes para o contexto desta tese. Os artigos selecionados foram submetidos à leitura completa para extrair informações sobre a sua natureza, sobre as técnicas utilizadas, sobre o seu foco e outras informações que pudessem contribuir para a tese. Esse número representa 51% dos artigos coletados.

A primeira análise, efetuada a partir da leitura dos artigos selecionados, segue a ordem cronológica de publicação. O artigo mais antigo selecionado foi publicado em 2004 e o mais recente, em 2013, ano corrente desta revisão.

A Figura 7 demonstra como está distribuída a quantidade de artigos selecionados para leitura ao longo dos anos.

Figura 7 – Artigos selecionados distribuídos por ano



Fonte: Elaborado pelo autor

Ao analisar a Figura 7, percebe-se que o número de publicações relacionadas à área vem aumentando a cada ano. Essa informação também pode ser constada ao analisar a Figura 4, que apresenta todos os artigos disponíveis sobre o tema.

Para cada palavra-chave que se utilizou como termo de busca, foram encontrados artigos relacionados com a tese. Além dos quatro termos utilizados, surge um quinto (foco) que pode ser referenciado como dicionário de sentimento. Nesse caso, o foco dos trabalhos está na criação de uma base de termos que possa auxiliar na classificação semântica, mais precisamente para a análise de sentimento. Formulou-se uma tabela com a quantidade total de artigos relevantes por termos de busca, a Tabela 1:

Tabela 1 – Totais de artigos divididos pelo seu foco

Termo / Foco	**Total**
Análise de sentimento	60
Análise de subjetividade	1
Classificação semântica	7
Mineração de opinião	13
Dicionários de sentimento	5

Fonte: Elaborado pelo autor

Pela leitura dos artigos, foi possível identificar as técnicas mais utilizadas para a classificação focadas na análise de sentimento. A Tabela 2 apresenta as técnicas ordenadas pelo seu uso nos artigos.

Tabela 2 - Técnicas utilizadas para a classificação.

Técnica	**Artigos que a utilizam**
SVM	15
POS *Tagging*	9
Clusterização	7
NaïveBayes	7
PMI	6
NER	4

Fonte: Elaborado pelo autor

Percebe-se que, para a etapa de classificação, são utilizadas técnicas de várias naturezas. O SVM (*Support Vector Machine*) que, segundo a revisão é a técnica mais utilizada, é conhecido como um

método de aprendizagem supervisionada. POS *Tagging*, que é a segunda técnica mais utilizada, baseia-se numa abordagem linguística.

A técnica de clusterização é classificada com uma tática para aprendizagem não supervisionada e é a terceira técnica mais utilizada. *Naïve Bayes* é uma técnica de aprendizagem supervisionada, da mesma forma que o SVM, e está empatada com a clusterização em se tratando de seu uso. A quinta técnica mais utilizada é a PMI (*Point Wise Mutual information*), uma abordagem de natureza estatística. A sexta, chamada reconhecimento de entidades nomeadas (NER – *Named Entity Recognition*), fundamenta-se em uma abordagem linguística.

Muitas outras técnicas foram empregadas, mas optou-se por trabalhar apenas com as seis mais utilizadas segundo a amostra recuperada. Além das técnicas, observou-se que muitos trabalhos utilizam como base para sua análise, dicionários, taxonomias, *corpus* anotados ou ontologias. Dos 85 artigos coletados, 45 utilizam alguma base de conhecimento para auxiliar na classificação.

A classificação semântica, mais precisamente, a análise de sentimento, pode ser aplicada a diferentes contextos. É possível observar alguns casos de aplicação a partir da leitura dos artigos selecionados.

- Análise da imagem de políticos (em período de eleição ou já eleitos);
- Opiniões sobre produtos ou serviços;
- Análise de citações de trabalhos científicos;
- Campanhas publicitárias e *marketing*;
- Inteligência competitiva;
- Detecção de crimes e de terrorismo;
- Identificação de situações críticas; entre outros.

Na seção a seguir, são apresentadas, com mais detalhes, a execução e as análises efetuadas a partir da segunda busca.

2.1.2.2 Busca 2 – Artigos relacionados com a temática

A segunda busca está centrada na identificação dos artigos que tratam tanto do tema – artigos eleitos na primeira busca – quanto da problemática. Percebe-se que a questão central do problema de pesquisa da tese encontra-se no reaproveitamento de raciocínios passados. Por conta disso, o termo de busca utilizado para selecionar os artigos desta fase foi *reasoning*.

Foram escolhidos 20 artigos que possuem o termo de busca em seu conteúdo. O problema dessa abordagem é que, mesmo que o termo de busca exista no conteúdo do documento, isso não garante que ele

esteja relacionado ao problema da tese. Por conta disso, foram selecionados apenas os artigos que estão diretamente relacionados com o tema em questão. Dos 7 artigos relacionados à problemática, 2 foram publicados em 2010, 2 em 2012 e 3 artigos publicados em 2013.

O primeiro artigo lido, desenvolvido por Wu et al. (2010), apresenta uma estratégia visual para auxiliar o analista em seu raciocínio. O grande foco desse trabalho está na forma de representar os sentimentos de uma massa de textos.

No trabalho de Huang e Qiu (2010), é apresentada uma técnica chamada SLN (*Semantic Link Network*) para a análise de citações. Nessa proposta, são construídos grafos semânticos, como aqueles em que se é possível raciocinar a partir da análise das relações dos nodos envolvidos. Nessa proposição, é viável utilizar regras e inferências sobre as relações como, por exemplo, raciocinar com fundamentação nas relações transitivas entre os nodos do grafo.

Cambria et al. (2012a) traz uma visão de raciocínio a partir de estruturas de conhecimento de domínio chamadas de ontologias. Na visão dos autores, deve-se utilizar para a análise de sentimento, a ideia do senso comum. São utilizadas várias ontologias de sentimento combinadas com uma base de sentenças, chamada *Open Mind Common Sense*, que possui dados coletados da Internet desde 2010. As ontologias permitem raciocínio com gênese na sua estrutura semântica, forma que o presente trabalho trata essa tarefa.

O trabalho apresentado por Cambria et al. (2012b), por sua vez, expõe o uso da mesma estratégia que o trabalho de 2012a, mas aplicada a outro domínio do conhecimento.

Na abordagem de Li e Tsai (2013), utilizou-se da lógica *Fuzzy* para tratar da classificação e da desambiguação da polarização. Para a aprendizagem, os autores usufruíram da técnica conhecida como Raciocínio Baseado em Casos (RBC). Essa tática permite que seja construída uma base de casos na qual, provindo do raciocínio, pode-se categorizar os novos documentos a partir da similaridade com os casos já conhecidos.

O trabalho de Pai (2013) apresenta uma etapa chamada Representação do Conhecimento e Raciocínio. Nessa etapa, é utilizada como base uma ontologia em que é possível fazer inferências (raciocínios) a partir da sua estrutura formal.

Kontopouloset al. (2013) propõe a utilização de uma ontologia de domínio para auxiliar na análise de sentimento. Para a parte de

raciocínio, é utilizada uma ferramenta de terceiros, chamada de *OpenDover*[2], que utiliza como base a ontologia proposta.

Após a leitura dos sete artigos retornados a partir da segunda busca, pôde-se observar que o tema raciocínio combinado com classificação semântica e a análise de sentimento está progredindo.

Percebe-se que o uso das ontologias como representação de conhecimento que pode ser raciocinado apareceu em cinco dos sete artigos selecionados, o que demonstra uma grande tendência na utilização desse recurso para a classificação e para a análise de sentimento. A próxima seção apresenta uma avaliação dos artigos seletos levando em consideração o problema de pesquisa.

2.1.2.3 Busca 3 – Artigos relacionados com o tema e o problema de pesquisa

A terceira e última busca da presente revisão sistemática tem como objetivo encontrar e verificar, a partir dos artigos retornados na Busca 2, quais trazem soluções para a problemática desta tese.

Verificou-se que nenhum artigo traz uma solução para a problemática proposta, que consiste na identificação de uma forma para armazenar e recuperar os raciocínios (ou inferências) anteriores com intuito de auxiliar nas novas classificações semânticas.

A abordagem que mais se aproxima da proposta é a construída por Li e Tsai (2013), a qual apresenta a utilização da técnica RBC. O uso desse método permite que, ao adicionar novos documentos (textos) em um processo de classificação (polarização), seja primeiramente verificado se não existem documentos similares já polarizados para que a classificação anteriormente já utilizada seja sugerida ao novo documento. Percebe-se que essa abordagem está muito focada no documento como aglomerado de termos, ao utilizar como base a similaridade dos termos presentes no documento. Contudo, não considera o seu domínio e o motivo pelo que o documento original faz parte da classificação que lhe foi dada.

O recurso das ontologias estava presente na grande maioria dos artigos selecionados como forma de dar contexto aos documentos que já fazem parte da base ou dos novos documentos, para se chegar a uma classificação mais precisa. As ontologias permitem que sejam aplicadas regras e inferências sobre as propriedades e as relações dos seus conceitos, ou seja, raciocinar sobre o seu domínio. Elas também

[2] Saiba mais em: http://opendover.nl

representam uma ótima alternativa para fazer uma classificação levando em consideração, no raciocínio, conceitos do domínio em questão.

O problema das abordagens dos trabalhos analisados é que não foi apresentado algum recurso que consiga reutilizar os raciocínios já efetuados no passado para novos casos, forçando que cada novo documento seja submetido a todo o processo. Dessa forma, percebe-se que não existem trabalhos, ponderando os artigos recuperados pela revisão sistemática, que tratem diretamente da problemática desta tese.

Tendo em vista as soluções apresentadas pelos artigos recuperados na segunda busca e analisando as suas características, formulou-se o seguinte pressuposto: Pode-se utilizar da técnica de raciocínio baseado em casos como forma para armazenar, recuperar, raciocinar e aprender a partir das inferências efetuadas numa ontologia de domínio, de modo que cada inferência executada seja tratada como um caso do modelo RBC.

A partir dessa afirmação, percebe-se que o uso da técnica de RBC, combinada com ontologias e métodos de classificação, pode compor um modelo de classificação que utilize raciocínios antecessores para novos documentos como uma nova forma para se obter a classificação semântica.

2.1.3 Etapa 3: Relatórios e divulgação

Esta revisão sistemática coletou, ao todo, 166 artigos que estavam disponíveis para *download* a partir da base *Web of Knowledge*. Desses, 85 foram selecionados como sendo relevantes para a tese.

A leitura dos artigos selecionados possibilitou analisar como as áreas de classificação semântica e de análise de sentimento estão organizadas. Percebe-se que muitas técnicas e modelos foram desenvolvidos para tratar do tema. A Tabela 3 apresenta a distribuição das técnicas utilizadas pelos anos dos artigos selecionados.

Tabela 3 - Técnicas distribuídas por ano

	2004	2006	2008	2009	2010	2011	2012	2013	Total
SVM		2	1	2	2	2	5	1	15
POS Tagging						4	5		9
Clusterização			1			4	2		7
Naïve Bayes		1	2	1			3		7
Linguística					2	1	3		6
PMI		1	1				3	1	6
NER						1	3		4
eWON							2	1	3
Fuzzy						1	1	1	3
PLN						1	1	1	3
Extração de collocations	1						1		2
FCA							1	1	2
Latent Dirichlet Allocation (LDA)							1	1	2
Lexicalized Hidden Markov Model							2		2
S-HAL (Sentiment Hyperspace Analogue to Language)							2		2
Agents							1		1
Algorítmo probabilistico CRF							1		1
Algorítmos genéticos (EWGA)			1						1
C4.5					1				1
Caminhamento randomico						1			1
Co-ocorrencia		1							1
FBS							1		1
GBS							1		1
Joint sentiment-topic							1		1
K-means					1				1
K-nearest neighbor			1						1
LM Classifier						1			1
LSA							1		1
Markov blanket model						1			1
Matriz de Palavras					1				1
Maximum Entropy				1					1
Método de triagulação							1		1
Método estatístico PCA							1		1
Método KDL					1				1
PageRank							1		1
PLSA						1			1
Rede Neural						1			1
Sentiment Probabilistic Latent Semantic Analysis							1		1
Weighted Structural Correspondence Learning						1			1
Wigner Function							1		1
Winnow Classifier			1						1

Fonte: Elaborado pelo autor

Como se pode perceber pelas técnicas utilizadas nos artigos e apresentadas na Tabela 3, muitas são classificadas como abordagens de aprendizagem supervisionada. Por conta disso, vários desses trabalhos utilizam uma base de sentenças já polarizadas como ‘partida fria’ da técnica.

O uso de dicionários e/ou bases de conhecimento também é uma prática bastante comum entre os modelos e os métodos de classificação semântica e de análise de sentimento. Dos 85 artigos coletados e selecionados, 45 utilizaram algum tipo de base de termos, taxonomia,

dicionários ou ontologias. Destaca-se os seguintes léxicos focados para a análise de sentimento:

- *SentiWordNet*;
- *WordNetAffect*; e
- *MicroWordNetOpinion*.

Dos léxicos citados, o mais utilizado é o *SentiWordNet*. Também foi encontrada uma ontologia focada em análise de sentimento chamada *EmotiNET*.

Buscou-se, entre os artigos selecionados, aqueles que, em seu modelo ou tema, apresentassem alguma abordagem relacionada ao uso de raciocínio. O resultado foi de 7 artigos, sendo que nenhum atendia ao problema inicial que era a possibilidade de armazenamento de um raciocínio de modo que este pudesse ser recuperado e utilizado para novas classificações. A Figura 8 apresenta os pontos positivos, em verde, e negativos, em vermelho, dos três grupos de abordagens, levando em consideração a problemática da tese.

Figura 8 – Características das abordagens resultantes da Busca 3

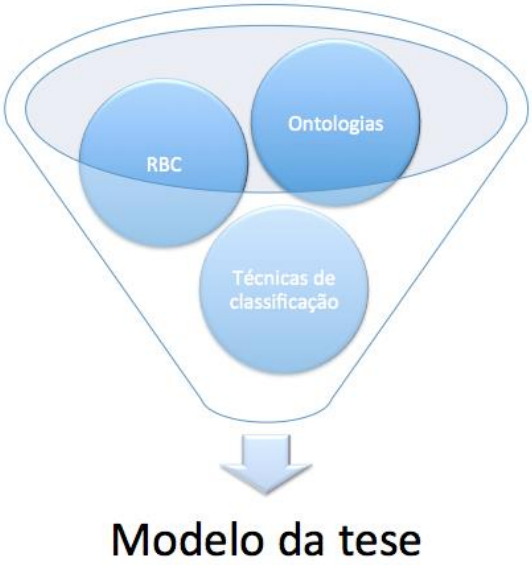
Fonte: Elaborado pelo autor

Percebe-se que ainda existe uma lacuna na área de classificação semântica e de análise de sentimento no que diz respeito ao

reaproveitamento de um raciocínio já concretizado, de modo a aproximar-se do que existe na aprendizagem humana.

A partir da revisão sistemática, pôde-se identificar elementos que devem estar presentes no modelo da tese. A Figura 9 apresenta essa proposta.

Figura 9 – Elementos do modelo da tese



Fonte: Elaborado pelo autor

Percebe-se que as técnicas de classificação podem ser utilizadas de maneira isolada e/ou combinada. Dessa forma, dependendo da natureza da classificação, pode-se ter um resultado melhor. Por conta disso, o modelo não deve estar focado em um método específico de classificação, podendo utilizar qualquer natureza.

O uso das ontologias permite que a abordagem seja sensível ao domínio em questão e também que seja possível inferir informações a partir das relações e das propriedades dos seus conceitos. Já o uso da técnica de RBC auxilia o modelo a tratar o armazenamento, a recuperação, a reutilização e a aprendizagem das inferências (raciocínios) executadas sobre as ontologias, tratando cada inferência como um caso do RBC.

2.2 CLASSIFICAÇÃO

Na visão de Kotsiantis, Zaharakis e Pintelas (2007), a classificação está diretamente ligada à aprendizagem de máquina, a qual pode ser supervisionada, quando utiliza rótulos já conhecidos para serem aplicados na classificação, ou não supervisionada, quando os rótulos são descobertos à medida que a técnica é aplicada.

A tarefa de classificação está presente numa grande parte das atividades efetuadas diariamente pelas pessoas. Computacionalmente, pode-se obter classificação a partir de abordagens estatísticas, via técnicas de aprendizagem de máquina, e por uso de redes neurais (MICHIE; SPIEGELHALTER; TAYLOR, 1994).

Segundo Monard e Baranauskas (2003), aprendizagem de máquina (ou *machine learning*) é uma área da inteligência artificial que tem como foco o desenvolvimento de técnicas computacionais para aprendizagem, bem como o desenvolvimento de aplicações capazes de adquirir conhecimento.

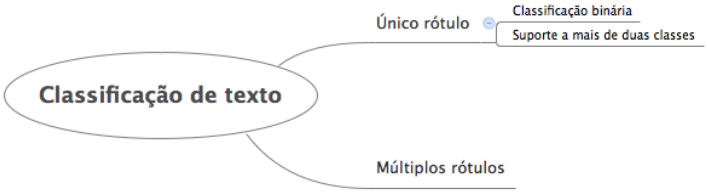
Para Von Wangenheim e Von Wangenheim (2003), pode-se definir o objetivo da classificação como classificar uma nova situação ou problema em um contexto específico. A seção a seguir foca na tarefa de classificação de texto, que é o tipo de classificação foco deste trabalho.

2.2.1 Classificação de texto

A tarefa de classificação de documentos ou textos faz parte do quotidiano da maioria das pessoas, seja na organização de matérias em um caderno, de documentos em uma pasta, ou ainda na forma como as pessoas organizam os seus arquivos no computador. Segundo Manning, Raghavan e Schütze (2009), a classificação consiste na organização (vinculação) de objetos a partir de um conjunto de classes.

Para Veeraselvi e Deepa (2013), classificação é uma abordagem geralmente dependente de treinamento de associação entre objetos e classes conhecidas. Normalmente, as categorias utilizadas na classificação estão diretamente ligadas a tópicos. Por conta disso, tal tarefa pode ser chamada de classificação de texto, categorização de texto ou classificação de tópicos (MANNING, RAGHAVAN e SCHÜTZE, 2009). A classificação de texto pode ser qualificada conforme a Figura 10.

Figura 10 – Tipos de classificação de texto



Fonte: GARCIA-CONSTANTINO, 2013

Na visão de Garcia-Constantino (2013), a classificação de texto pode ser categorizada de duas formas: (1) como sendo baseada em um único rótulo, ou seja, permite ter apenas uma classe vinculada ao objeto em questão, ou (2) baseada em múltiplos rótulos, admite acoplar mais de uma divisão ao mesmo objeto.

O autor ainda afirma que, quando se tem uma classificação baseada em um único rótulo, pode-se qualificá-la como sendo uma classificação binária, ou seja, suporta apenas dois classificadores, como, por exemplo, a classificação positivo/negativo, ou então pode suportar mais de duas classes como possibilidades de classificação.

Quando se trabalha com uma quantidade pequena de textos e de documentos, a classificação pode ser efetuada de maneira manual. Todavia, quando se tem como cenário um conjunto maior de documentos, essa tarefa passa a ser mais custosa. Segundo Buche, Chandak e Zadgaonkar (2013), para tratar o volume massivo de dados disponíveis na *Web* e nas organizações, devem-se utilizar o auxílio de abordagens computacionais.

Para Wang e Domeniconi (2008), a categorização de textos representa um desafio para as áreas da mineração de dados e da aprendizagem de máquinas, subáreas da computação que procuram desenvolver métodos computacionais para auxiliar na classificação dos textos digitais. Para atacar esse problema, Prabowo e Thelwall (2009) afirmam que o uso da classificação automática de textos e de documentos é indicado. Os autores apresentam alguns tipos de classificação:

- **Classificação baseada em regras**: nesse modelo são construídas regras que auxiliam na categorização do texto a partir das classes levantadas.
- **Classificação baseada em estatística**: utiliza métodos de probabilidade e de estatística para verificar informações como

> frequência e relação de coocorrência entre termos para chegar a uma classificação do texto em questão.

Além das abordagens apresentadas pelos autores, pode-se apresentar a classificação baseada na linguística, que utiliza como base a classe gramatical dos termos e, em alguns casos, a sua semântica para auxiliar na classificação.

Uma das áreas das aplicações da tarefa de classificação é a análise de sentimento. Para Manning, Raghavan e Schütze (2009), a classificação de sentimento é um importante campo de aplicação da qualificação de texto. Nas Seções 2.3 e 2.4, são apresentados mais elementos sobre a área da análise de sentimento.

2.3 SENTIMENTO

O termo sentimento é utilizado e estudado por muitas áreas, como por exemplo: a psicologia, as ciências biológicas, a antropologia, a filosofia, a ciência da computação, entre outras. Segundo Ferreira (2009), pelo do dicionário Aurélio:

> Ato ou efeito de sentir. / Aptidão para sentir; sensibilidade. / Sensação íntima, afeto: os sentimentos de um pai. / Conhecimento imediato; intuição: tem o sentimento de seu valor. / Dor, mágoa, desgosto. / Qualidades ou tendências morais: estar animado de bons sentimentos. / Pêsames: aceite meus sentimentos.

Segundo Franz (2003), a base da psicologia para conceitualizar o sentimento veio dos estudos dos filósofos Platão, Aristóteles, Descartes, Spinoza, Hume e Kant. Sentimentos como designação das emoções, simpatias e suscetibilidade surgem em 1771, juntamente com a utilização de algumas palavras para este fim, como, por exemplo, interessante, entediado, tédio, saudade, acanhamento, apatia etc.

Sobre o significado de sentimento, Thums (1999), apoiado nos estudos de Agnes Heller[3], afirma que "sentir significa estar implicado em algo!", que pode ser outra pessoa, uma ideia, um produto, um problema, uma situação.

Franz (2003) afirma que os sentimentos não são apenas de natureza pessoal, eles também refletem fenômenos de cunho histórico e

[3]Agnes Heller nasceu em Budapeste em 1929, foi professora de sociologia na Universidade de Trobe, na Austrália. Possui publicações na área da sociologia e estudos sobre os sentimentos (THUMS, 1999).

universal, podendo ser comuns e coletivos. Jung[4] trabalhou o conceito de sentimento como objeto de estudo. Ele descobriu o papel dos sentimentos partindo da observação de seus experimentos de associação, nos quais se deparou com relações afetivas puras (“sim”, “ruim”, “gosta”, entre outras) (FRANZ, 2003).

Para Thums (1999), o sentimento não representa apenas uma informação, mas pode também ser entendido como motivação, ou seja, está diretamente ligado à tomada de decisão. Segundo Franz (2003), “A função sentimento é o processo psicológico que avalia. Por seu intermédio, apreciamos uma situação, pessoa, objeto ou momento em termos de valores.”.

Sobre a avaliação cognitiva do sentimento, Fialho (2011) explica que inicialmente acontece uma avaliação primária, não diferenciada, do tipo positiva ou negativa, e em um segundo momento, chega-se ao sentimento/emoção mais preciso como, por exemplo, vergonha, gratidão, satisfação etc. Franz (2003, p.150) afirma “Os sentimentos como conteúdo da psique podem ser qualificados por um sinal positivo (+) ou negativo (-).”.

Segundo Thums (1999), os sentimentos são experiências subjetivas que devem ser analisadas pela perspectiva de quem as vive. Para Mejova (2011), definir sentimento não é uma tarefa simples, sabe-se que o seu conceito está diretamente ligado com os referentes à subjetividade e opinião. Percebe-se que o sentimento descrito pelo autor está diretamente ligado à definição de opinião.

Na seção a seguir são apresentados mais detalhes sobre a relação da opinião com o sentimento.

2.3.1 Opinião como forma de sentimento

Toda opinião possui, de forma implícita ou explícita, um sentimento agregado, ou seja, um sentido positivo ou negativo sobre um evento, produto ou situação (KIM; HOVY, 2004). Segundo Pang e Lee (2008), para a área de coleta de dados, a opinião sempre foi uma importante informação a ser considerada, já que ela apresenta o que as pessoas pensam, podendo auxiliar na tomada de decisão.

As opiniões desempenham um papel fundamental para quase todas as ações humanas, desde a forma de pensar o que fazer e como agir (PADMAJA; FATIMA, 2013). Para Esuli (2008), elas apresentam

[4] Carl Gustav Jung (1875 — 1961), psiquiatra e psicoterapeuta suíço, construiu a psicologia analítica. Ele definiu e implementou as definições de arquétipo, personalidade extrovertida e introvertida e inconsciente coletivo (LACHMAN, 2010).

dados sobre como a realidade é percebida por outras pessoas, ou seja, como são os seus sentimentos a respeito de determinada coisa.

Usam-se opiniões para expressar pontos de vista¸ de modo que os pensamentos de outras pessoas podem ser úteis para amparar o veredito. As organizações também fazem uso desse tipo de ferramenta por meio das opiniões do seu público-alvo, as quais podem ser um importante direcionador. Indústrias realizam pesquisas de mercado coletando as avaliações das pessoas sobre os seus produtos e os de seus concorrentes a fim de compará-los e assim planejar suas estratégias de mercado (ESULI, 2008).

Segundo Pak e Paroubek (2010), existem muitas questões que as organizações gostariam de responder baseadas na opinião dos seus clientes, como, por exemplo:

- O que as pessoas acham sobre o nosso produto (serviço, organização etc.)?
- Nossa imagem está positiva ou negativa?
- O que poderia ser melhor no nosso produto na visão dos nossos clientes?

Para dar suporte às análises apresentadas no parágrafo anterior, o campo da computação, mais precisamente dos sistemas de apoio à decisão, aproveitam técnicas de processamento de linguagem natural e estatística com o intuito de facilitar o processamento de sentenças. Para que se possa utilizar o supracitado mecanismo computacional, é necessário armazenar e recuperar essas informações em um repositório.

A próxima seção apresenta os possíveis repositórios, acessíveis por recursos computacionais, nos quais é possível recuperar e processar informações relacionadas a opiniões e sentimentos de usuários.

2.3.2 Repositórios de dados de sentimentos e opiniões

Para que as informações sejam processadas e utilizadas no percurso da tomada de decisão, faz-se necessário que estejam disponíveis de forma codificada em uma mídia ou em um canal. É sabido que as fontes de dados não-estruturados podem conter uma grande quantidade de informações e de conhecimentos implícitos (AGARWAL et al., 2015).

A coleta de opiniões inicialmente era efetuada por entrevistas e/ou submissão de questionários a uma amostra de pessoas. Com a evolução da Internet, tornou-se possível que usuários registrem ou consultem avaliações sobre produtos, serviços etc. (PANG e LEE, 2008).

Segundo Tang, Tan e Chieng (2009), revistas e jornais especializados já faziam revisões sobre produtos e serviços a fim de obter um parecer que facilitasse a tomada de decisão dos seus leitores. Com o advento da Internet, essas mídias foram migrando, aos poucos, para a disponibilização do seu conteúdo também no formato digital. A *Web* passa, nesse contexto, a ser um importante canal para coletar e expressar pensamentos e sentimentos. Para Pang e Lee (2008), 81% dos usuários da *Web* fazem pesquisa sobre um produto antes de adquiri-lo. E desses 81%, 20% fazem esse tipo de busca diariamente.

A *Web* evoluiu para o fenômeno chamado *Web 2.0*, que tem como principal característica a mudança do papel do usuário de um simples consumidor de dados e de informação para também produtor desses dados. Isso só foi possível com a criação de sistemas e plataformas em que o usuário não precisa conhecer conceitos de programação ou outros recursos mais avançados da computação para publicar conteúdo, pois esses sistemas fazem a interface, permitindo que, de maneira simples, os dados sejam cadastrados e disponibilizados na forma de página da *Web*.

Para O'Reilly (2007), o termo *Web 2.0* cunhado por ele e pela *MediaLive International*[5] representa uma mudança no papel de seus usuários, o que possibilitou uma explosão de novos conteúdos e serviços. Segundo Torres (2009), o termo *Web 2.0* não se trata de uma tecnologia específica, mas sim de um fenômeno comportamental na Internet.

Segundo Zabin e Jefferies (2008), com o grande aumento das plataformas da *Web 2.0*, como *blogs*, fóruns de discussão, redes de colaboração e vários outros tipos de mídias sociais, os seus usuários e as organizações percebem que muitas informações, positivas ou negativas, sobre produtos e serviços, são trocadas nesse meio e que, cada vez mais, esse tipo de informação é utilizada para a tomada de decisão.

Neste trabalho, objetiva-se operar com repositórios de opiniões e sentimentos disponíveis nas plataformas da *Web 2.0*. As próximas subseções apresentam mais detalhes sobre algumas dessas plataformas.

2.3.2.1 *Blog*

Um *blog* representa uma plataforma computacional que possibilita, aos usuários, adicionar conteúdo textual ou multimídia de maneira simples na *Web*. Segundo Godbole, Srinivasaiah e Skiena

[5] A *MediaLive International* é uma organização que produz, gerencia e promove eventos para a indústria de tecnologia da informação. Veja mais em http://10times.com/organizers/media-live.

(2007), tanto jornais como *blogs* expressam opiniões, novidades sobre pessoas, lugares, coisas e eventos recentes. Para esses autores, o uso de *blogs* como fonte de informação é tão rico e expressivo quanto o uso de um jornal.

Segundo O'Reilly (2007), um *blog* é uma página pessoal no formato de diário que, se for analisada a partir da dimensão tempo, pode trazer informações importantes para pessoas e organizações. Segundo Telles (2011), existem aproximadamente 152 milhões de *blogs* disponíveis na Internet.

Para Tang, Tan e Cheng (2010), os *blogs* permitem que pessoas possam compartilhar suas visões, opiniões e sentimentos, tornando bastante comum a construção de *blogs* especializados, os quais direcionam o conteúdo a um público-alvo. Percebe-se que esse modelo de plataforma disponibilizou um caminho para que as pessoas que não conheciam informática pudessem compartilhar com os demais usuários da *Web* seus textos, imagens, áudios e vídeos.

Muito frequentemente encontrarmos *blogs* temáticos, nos quais os seus autores apresentam relatos sobre a utilização de produtos e dicas para compra de novos equipamentos. Essas páginas representam uma importante fonte de informação para os consumidores que, geralmente, as consultam antes de fazer um compra ou contratar um serviço. Um exemplo de *blog* temático é o Reclame Aqui[6]. Nele é possível encontrar opiniões sobre serviços e produtos e, devido ao grande número de acessos, tanto empresas quanto usuários pessoa física frequentam a plataforma.

Os *blogs* são utilizados como fonte de dados para analisar sentimentos e opiniões, como é possível observar nos trabalhos de:Ku, Liang e Chen (2006); Godbole, Srinivasaiah e Skiena (2007); Tang, Tan e Cheng (2010), entre outros.

2.3.2.2 *Microblogging*

Os *blogs* representam uma importante ferramenta para que os usuários da Internet possam publicar seus textos de maneira simples e rápida. Os *microbloggings* surgem com uma proposta um pouco diferente, cujo foco não é publicar um texto, mas sim uma mensagem, de maneira simples e rápida.

Para Pak e Paroubek (2010), *microbloggings* são plataformas usadas por diferentes pessoas para expressar opiniões sobre diferentes tópicos ou falar sobre eventos de sua vida. Na visão de Narr, Hulfenhaus

[6]Acesse o blog Reclame Aqui pelo endereço: http://www.reclameaqui.com.br/.

e Albayrak (2012), são plataformas em que as pessoas podem trocar mensagens curtas sobre qualquer assunto.

Os *microbloggings* representam hoje uma das mais populares formas de comunicação para a maioria dos usuários, sendo que milhões de mensagens são trocadas diariamente (PAK; PAROUBEK, 2010).

Segundo Aisopos et al. (2012), a plataforma de *microblogging* mais popular e utilizada entre os usuários da *Web* é o Twitter[7], o qual possui mais de 180 milhões de usuários que publicam mais de 1 bilhão de mensagens por semana. Para Go, Bhayani e Huang (2009), o Twitter é um serviço popular de *microblogging* em que o usuário pode publicar mensagens sobre o seu estado atual. Ademais, essas mensagens podem conter opiniões sobre diferentes tópicos.

Narr, Hulfenhaus e Albayrak (2012) explicam que as mensagens trocadas pelo Twitter são chamadas de *tweets* e devem possuir no máximo 140 caracteres, o que obriga os seus usuários a serem diretos em suas mensagens. Segundo Hu et al. (2013), *microbloggings* como o Twitter permitem ao ser humano expressar de maneira fácil seus sentimentos, opiniões, notícias e eventos de sua vida. Por conta disso, essas plataformas vêm sendo muito utilizadas para a coleta e a análise de informações estratégicas, visando a tomada de decisão da organização.

Segundo Deng et al. (2013), as informações postadas nos *microbloggings* são utilizadas para identificar como está a imagem, seja ela positiva ou negativa, de pessoas, produtos, marcas e organizações. Os autores afirmam ainda que esse tipo de conteúdo é muito utilizado para análise de popularidade de candidatos nos Estados Unidos e em alguns países da Europa durante o período de campanha eleitoral.

Confirmando as afirmações efetuadas por Deng et al. (2013), o trabalho de Hu, Wang e Kambhampati (2013) apresenta um modelo para analisar eventos a partir dos *tweets* publicados durante um período específico. O estudo de caso do referido trabalho foi baseado nas mensagens geradas durante o debate dos candidatos à presidência dos Estados Unidos em 2012 e apresentou resultados positivos.

Diante do que foi apresentado nesta seção, percebe-se que os *microbloggings*, principalmente o Twitter, têm sido muito utilizados como fonte de informação para análise de opinião e de sentimento. Isso se dá pela quantidade de usuários desse tipo de plataforma e pela facilidade de comunicação que ela proporciona.

2.3.2.3 Redes sociais

[7] Acesse o Twitter em https://twitter.com/.

Seguindo o movimento da *Web 2.0* com plataformas que tornaram mais fácil publicar conteúdo na *Web*, tais como, os *blogs* e *microblogging*, surgem as redes sociais, que têm como foco a integração de pessoas na *Web* a partir de plataformas computacionais. Para Boyd e Ellison (2008), redes sociais são serviços, baseados na *Web*, que permitem aos indivíduos construir perfis públicos dentro de um sistema limitado, podendo articular listas de amigos (outros usuários) para interações.

Segundo Adolpho (2011), as redes sociais e suas comunidades funcionam como canais nos quais os usuários expressam seu estilo de vida para os amigos de sua rede. Na visão de Telles (2011), as redes sociais, ou *sites* de relacionamento, são ambientes que têm como objetivo reunir pessoas, possibilitando que os usuários troquem mensagens, fotos, vídeos e se relacionem.

Para Simranjit, Nikunj e Nishnt (2012), as redes sociais na *Web* nasceram com foco na integração entre as pessoas, mas, ao longo da sua utilização, as organizações e o meio acadêmico voltaram seu olhar para os dados e para as informações geradas a partir da integração entre os usuários nesses ambientes.

Com a nova dinâmica proveniente das relações sociais a partir das redes sociais, a interação e a comunicação tornam-se parte do processo de criação do conhecimento dos seus membros (BALANCIERI, 2010).

As redes sociais disponibilizam canais, na forma de comunidades, que agrupam membros interessados por um tema em comum. Os conteúdos gerados nessas comunidades podem trazer informações bastante válidas para a tomada de decisão. Segundo Chandran e Murugappan (2012), os *sites* de relacionamento apresentam uma série de informações, visões e opiniões sobre produtos, serviços, eventos, pessoas etc. Cada postagem efetuada pode ser utilizada por ferramentas e por técnicas para auxiliar na tomada de decisão da organização, afirma o autor.

Para que as entidades possam extrair e processar os dados referentes a opiniões e sentimentos existentes nas mídias sociais (*blogs*, *microblogging*, redes sociais, etc.), faz-se necessária a utilização de técnicas e ferramentas. A área de análise de sentimento tem como função a obtenção de tais necessidades. Na seção seguinte, são apresentados os conceitos relacionados à análise de sentimento bem como as suas principais técnicas, métodos e ferramentas.

2.4 ANÁLISE DE SENTIMENTO

O desafio de classificar e analisar sentenças com o propósito de identificar opiniões e sentimentos não é recente na computação. Podem-se encontrar trabalhos como os apresentados por Kim e Hovy (2004), em que Hatzivassiloglou e McKeown (1997) já se preocupavam com a identificação de opiniões, ou trabalhos como os de Wiebe et al. (2002) e Riloff et al. (2003), mais focados em identificar a subjetividade inclusa nas opiniões e sentimentos das sentenças.

Verifica-se que esses desafios persistem até os dias atuais, em que as fontes de consumo de dados e de informações evoluíram em paralelo aos movimentos da *Web*. Anteriormente, eram utilizados como fonte para essas análises, textos publicados por *sites*. Posteriormente, os jornais foram ingressando na Internet e publicando notícias por esse meio. Assim, passou-se a fazer uso também das notícias publicadas em jornais *online*. Com o surgimento das mídias sociais (*blogs*, *microbloggings*, redes sociais etc.), nas quais os usuários podem publicar, de maneira rápida e fácil, o conteúdo, formou-se um importante canal que pode ser utilizado como fonte.

As áreas de análise de sentimento e de mineração de opinião surgem com o objetivo de atuar na identificação de soluções computacionais para os desafios de identificar, classificar e analisar sentimentos e opiniões de sentenças.

Segundo Pang e Lee (2008), a análise de sentimento tem como objetivo polarizar (classificar) uma sentença como positiva ou negativa. Para Narayanam, Liu e Choudhary (2009), a análise de sentimento também pode ser chamada, em muitos casos, de mineração de opinião – técnica, na qual se tem como foco a classificação de um documento ou sentença como positiva ou negativa. Na visão de Liu (2010a), a análise de sentimento ou mineração de opinião é a área da computação que estuda as opiniões, os sentimentos e as emoções expressas em texto.

Para Guerra et al. (2011), análise de sentimento e a mineração de opinião são áreas que pesquisam, principalmente, a extração de opinião de cenários bem controlados e não dinâmicos. Os autores afirmam que o cenário deve ser controlado pelo fato de a análise de sentimento e de opinião estar diretamente ligada a um contexto, sendo assim, uma informação bastante subjetiva de quem a produziu.

Nesta seção, objetivou-se apresentar os principais conceitos a cerca da área de análise de sentimento, como ela está dividida e quais os seus principais desafios e técnicas. Na próxima seção, expõe-se um breve resumo a respeito dos fatos históricos ligados à área de interesse.

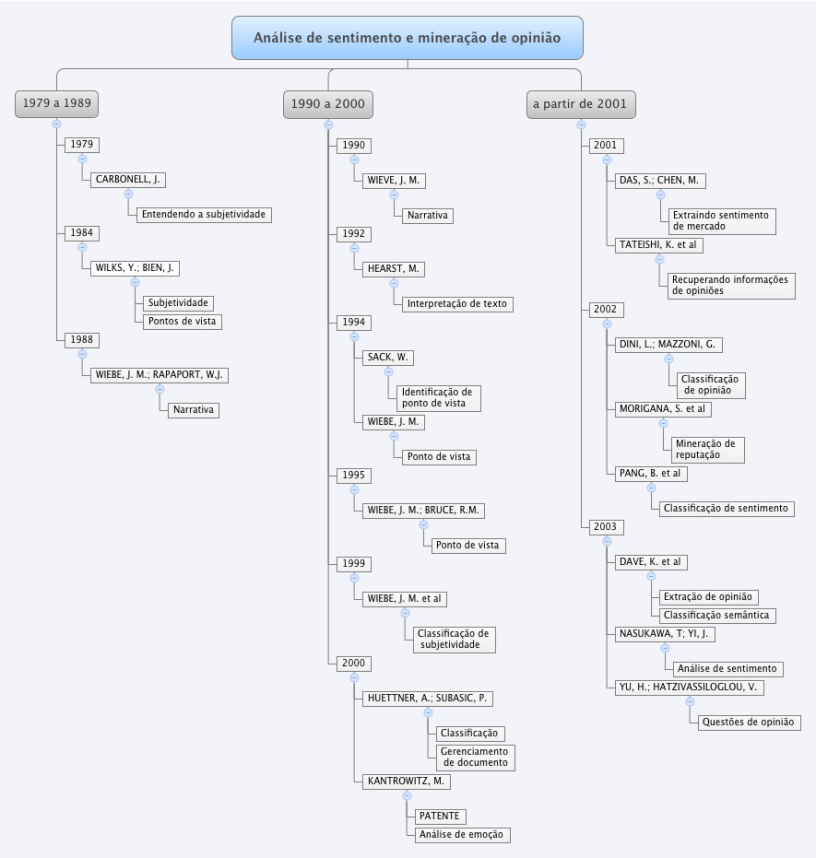
2.4.1 Histórico

Nos últimos anos, as áreas de análise de sentimento e de mineração de opinião têm produzido um número crescente de trabalhos e pesquisas no meio acadêmico. No entanto, pesquisas relacionadas e inspiradoras para a área já vinham sendo construídas desde o final da década de 70 (PANG; LEE, 2008; SERRANO-GUERRERO et al., 2015).

Segundo Pang e Lee (2008), a partir de 2001, os pesquisadores reconhecem os problemas relacionados à área de análise de sentimento e de mineração de opinião juntamente com a sua importância, gerando uma série de pesquisas e artigos.

A Figura 11 apresenta, com base nos trabalhos informados no artigo de Pang e Lee (2008), uma visão geral sobre os primeiros passos da área da análise de sentimento e de mineração de opinião.

Figura 11 – Início da área de análise de sentimento



Fonte: Adaptado de PANG; LEE, 2008.

Com base nesses dados, percebe-se que os estudos inicias eram focados principalmente na identificação da subjetividade e do ponto de vista das pessoas em textos. Em paralelo, também eram pesquisados métodos para analisar narrativas. Segundo Wiebe et al. (2004), subjetividade em linguagem natural refere-se ao aspecto da língua utilizado para expressar opiniões, avaliações e especulações.

Somente no final da década de 90 é que se iniciam os estudos da classificação da subjetividade identificada nos textos. No ano de 2000, é registrada a primeira patente relacionada à área de análise de emoções (PANG; LEE, 2008). Conforme apresenta-se na Figura 11, é possível visualizar que a área ganhou destaque e foco na análise das opiniões e de sentimentos a partir de 2001.

Para Westerski (2012), mesmo que os primeiros trabalhos na área de análise de sentimento tenham surgido entre as décadas de 80 e 90, tentando identificar a subjetividade, a área ganhou visibilidade com o advento da *Web 2.0*. Segundo Pang e Lee (2008), os motivos que impulsionaram o desenvolvimento da área de análise de sentimento e mineração de opinião a partir de 2001 foram:

(1) a evolução das técnicas e dos métodos de processamento de linguagem natural e de recuperação de informação;
(2) o desenvolvimento da *Web* e dos repositórios de dados disponíveis nela; e
(3) as demandas comerciais e de *marketing* por análises dessa natureza.

A área de análise de sentimento faz referência ao seu objetivo de estudo, identificar o sentimento (se positivo ou negativo) de um texto, com muitos nomes distintos. A seção, a seguir, apresenta mais detalhes sobre a terminologia da área.

2.4.2 Terminologia

Segundo Pang e Lee (2008), muitos nomes são dados para essa atividade como, por exemplo: mineração de opinião (em inglês, *opinion mining*), análise de sentimento (em inglês, *sentiment analysis*) ou análise de subjetividade (em inglês, *subjectivity analysis*). Os autores ainda afirmam que também são utilizados os termos mineração de revisão (em inglês, *review mining*) e extração de avaliação (em inglês, *appraisal extraction*), mas esses casos geralmente estão mais relacionados com a área da computação afetiva.

Sobre a definição de análise de subjetividade, segundo Wiebe (2000), é uma forma de separar opiniões individuais de fatos. Segundo Banea, Mihalcea e Wiebe (2008), subjetividade, no contexto do trabalho dos autores, são as emoções, opiniões e os sentimentos expressos em textos. A análise de subjetividade tem como objetivo identificar e classificar essas características. Para Zhou e Chaovalit (2008), a análise ou mineração de subjetividade tem como função identificar os termos subjetivos. Na visão dos autores, essa pode ser uma etapa de pré-processamento para a mineração de opinião ou para a análise de sentimento. Para Padmaja e Fatima (2013), a análise de subjetividade utiliza-se de vários métodos e técnicas originadas nas áreas de recuperação da informação, inteligência artificial e processamento de linguagem natural.

Na visão de Pang e Lee (2008), os termos mineração de opinião e análise de sentimento surgiram na literatura paralelamente. Para alguns

autores, são conceitos distintos. Segundo Liu (2010a), tanto a análise de sentimento quanto a mineração de opinião tem como função a identificação de emoções e sentimentos em textos.

Para Westerski (2012), mineração de opinião é, muitas vezes, referida como análise de sentimento na literatura, sendo que o seu foco é aproveitar a grande quantidade de conteúdo publicado pelos usuários a partir de recursos da *Web 2.0* para extrair conhecimentos sobre opinião, para apoio às futuras decisões.

Na visão de Padmaja e Fatima (2013), a mineração de opinião está mais focada na identificação e extração da subjetividade de documentos textuais, já a análise de sentimento tenta apresentar o sentimento (polarização) da sentença, ou seja, se ela é positiva ou negativa. Segundo Abbasi, Chen e Salem (2008), a análise de sentimento está concentrada em analisar textos contendo opiniões e emoções, já a área de classificação de sentimento tem como foco determinar se um texto é subjetivo ou objetivo, e, quando subjetivo, se positivo ou negativo.

O presente trabalho utiliza como termo-chave a análise de sentimento. O motivo para tal escolha é que o objetivo deste trabalho não se limita à identificação ou extração da subjetividade existente em textos, mas preocupa-se sim, com a classificação das sentenças.

2.4.3 Métodos e técnicas

Nesta seção, são apresentadas as seis técnicas que foram mais utilizadas pelos artigos analisados durante a fase da revisão sistemática. As técnicas selecionadas foram: SVM, POS *Tagging*, Clusterização, *NaïveBayes*, PMI e reconhecimento de entidades nomeadas, *deep learning* e SVD.

2.4.3.1 Support Vector Machine (SVM)

O *Support Vector Machine* (SVM), que em português pode ser traduzido por Máquina de Vetores de Suporte, foi proposto por Vladimir N. Vapnik, inicialmente, como classificadores lineares e depois como classificadores não lineares (HAYKIN, 2001).

Segundo Rajper et al. (2012), SVM é uma técnica popular de classificação baseada em aprendizagem de máquina. O objetivo dessa técnica é encontrar o limite para decidir entre a classificação em duas classes, utilizando treinamento de dados. O trabalho de Mullen e Collier (2004) apresenta o uso de *Support Vector Machines* (SVM) para auxiliar na tarefa de classificação e análise.

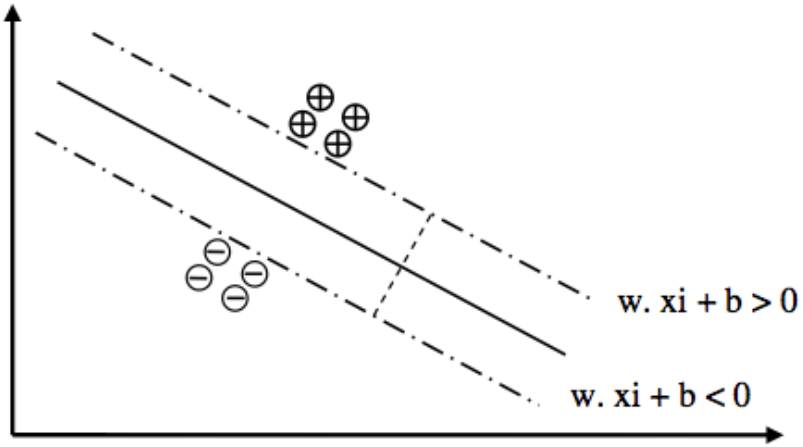
Para se utilizar o SVM no contexto da análise de sentimento, Padmaja e Fatima (2013) apresentam a seguinte explicação: tendo um conjunto de categorias (positivo e negativo), C ={+1, -1}, e dois conjuntos de documentos pré-classificados para treinamento, um conjunto positivo, representado pela Equação 1, e um conjunto negativo, representado pela Equação 2.

$$TR^{+} = \sum_{i=1}^{n} (di, +1) \quad (1)$$

$$TR^{-} = \sum_{i=1}^{n} (di, -1) \quad (2)$$

O SVM encontra o hiperplano que separa o conjunto das duas classes, demonstrando a margem de separação. A Figura 12 expõe essa separação.

Figura 12 – Separação efetuada pelo método SVM



Fonte: Padmaja e Fatima (2013)

Cada conjunto de documentos pré-treinados é convertido em vetores reais, *xi* consiste em um conjunto de recursos significantes a documentos associados, *di*, onde b é um termo qualquer.

Para Manning, Raghavan e Schütze (2009), o SVM além de definir o limiar de separação, constrói o que é chamado de margem de separação, que, na Figura 12, é representada pelas linhas pontilhadas.

Segundo Haykin (2001), o objetivo do SVM é encontrar o hiperplano particular para o qual a margem (pontilhada) seja a máxima.

Essa informação é importante, pois define os limites no plano da classificação dos elementos.

A técnica de SVM é muito utilizada no contexto de classificações de documentos e de textos, e não é diferente na análise de sentimento. Pode-se fazer essa afirmação pelos dados extraídos durante o processo de revisão sistemática deste trabalho, que apresentou, dentre os trabalhos selecionados para a leitura, essa técnica como a mais utilizada.

Na seção seguinte, são expostos mais detalhes sobre a técnica de POS *Tagging*.

2.4.3.2 Part-Of-Speech Tagging

Quando se pretende fazer uma análise detalhada sobre um texto, geralmente ele é dividido em sentenças e depois cada palavra é analisada, de modo que é feita uma análise linguística. Para isso, utiliza-se o *part-of-speech* (POS) de cada palavra (WEISS et al., 2005).

Segundo Manning e Schütze (1999), o POS *Tagging* é uma tarefa da área de processamento de linguagem natural que tem como objetivo entender a linguagem, de modo que as classes gramaticais das palavras sejam explicitadas.

Por meio da identificação da classe gramatical de cada termo de uma sentença, é possível definir padrões linguísticos para auxiliar no processamento da informação presente nela. Pode-se utilizar POS *Tagging* em diferentes situações, como, por exemplo: recuperação de informação, análise de uma sentença, classificação de um texto como positivo ou negativo, auxílio num processo de desambiguação, entre outros (MANNING; SCHÜTZE, 1999).

Como exemplo para a utilização do POS *Tagger*, apresenta-se a seguinte sentença: *"Most large cities in the US had morning and afternoon newspapers."*. A Figura 13 apresenta um exemplo de saída obtida pela aplicação do POS *Tagger*.

Figura 13 – Exemplo do texto processado pelo POS *Tagger*

```
Most ==> Adjective, superlative
large ==> Adjective
cities ==> Noun, plural
in ==> Preposition or subordinating conjuntion
the ==> Determiner
US ==> Proper noun, singular
had ==> Verb, past tense
morning ==> Noun, singular or mass
and ==> Coordinating conjunction
afternoon ==> Noun, singular or mass
newspapers ==> Noun, plural
```

Fonte: Elaborado pelo autor

Observando a Figura 13, pode-se verificar que, para cada palavra, foi identificada a classe gramatical correspondente, o que permite que a linguagem natural expressa no texto possa ser mais facilmente processada, de modo a extrair informações importantes para o contexto de aplicação.

A seção a seguir apresenta a técnica de aprendizagem de máquina chamada de clusterização ou agrupamento.

2.4.3.3 Clusterização

Clusterização é uma técnica de aprendizagem de máquina classificada como não supervisionada, ou seja, não possui nenhuma forma de treinamento dos dados para gerar uma nova classificação (KONCHADY, 2006).

Segundo Carpineto et al. (2009), a clusterização é utilizada constantemente em problemas de classificação e para recuperação de informação, de modo que os resultados são organizados em grupos (*clusters*). Ainda na visão dos autores, clusterização pode ser caracterizada como um processo de descoberta de subconjuntos de objetos, no qual os itens de um grupo possuem características em comum.

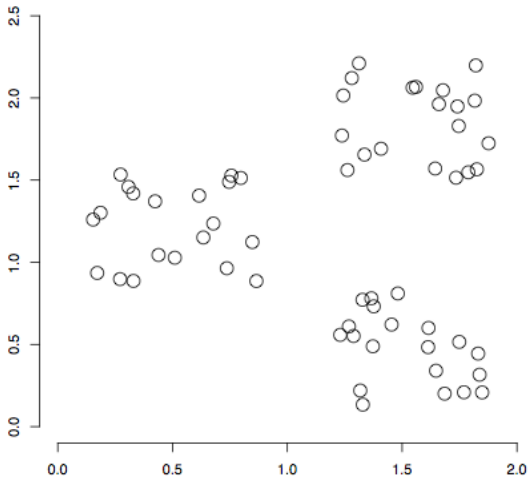
Beil, Ester e Xu (2002) explicam que a maioria dos algoritmos de clusterização de texto trabalha utilizando a abordagem chamada de modelo de vetor de espaço, na qual cada documento, *d*, é representado por um vetor de frequências de termos: $d = (tf_1, \dots, tf_n)$. Os vetores de termos são normalizados para medir a similaridade entre dois documentos, *d1* e *d2*, que representam os vetores de frequência de termos. A similaridade pode ser obtida por meio da equação do cosseno que mede o ângulo entre os dois vetores, conforme representado na Equação 3, a seguir.

$$similaridade\ (d_1, d_2) = \frac{(d_1 \circ d_2)}{\| d_1 \| . \| d_2 \|} \quad (3)$$

Nessa equação, $(d_1 \circ d_2)$ apresenta o produto dos dois vetores e $\|$ representa o tamanho do vetor em questão.

Para Manning, Raghavan e Schütze (2009), a clusterização pode ser utilizada em diferentes situações como, por exemplo, agrupar os resultados a partir de uma busca textual, organizar documentos com referência ao seu conteúdo, criar conjuntos de palavras com base em sua coocorrência etc. A Figura 14 apresenta um exemplo dessa tarefa.

Figura 14 – Exemplo de clusterização



Fonte: Mannning, Raghavan e Schütze (2009)

A seção a seguir trata de mais uma técnica, conhecida como *Naïve Bayes*.

2.4.3.4 Naïve Bayes

Antes de apresentar os conceitos relacionados com *Naïve Bayes* (NB), é importante conhecer a estatística bayesiana. Segundo Manning e Schütze (1999), as abordagens bayesianas são fundamentadas em estatísticas e muito utilizadas para auxiliar no processamento de linguagem natural. *Naïve Bayes* é uma técnica de aprendizado de máquina supervisionado, muito utilizada na tarefa de classificação, a partir de um conjunto de documentos inicias utilizados para treinamento (MANNING; RAGHAVAN; SCHÜTZE, 2009).

Segundo Chen et al. (2009), o uso do *Naïve Bayes* é bastante simples e apresenta um resultado muito bom na tarefa de classificação. Na visão de Rish (2001), classificadores bayesianos atribuem a classe mais provável para um objeto a partir do seu vetor de características.

Manning, Rachavan e Schütze (2009) afirmam que *Naïve Bayes* é baseado em métodos probabilísticos, nos quais se levanta a probabilidade de um elemento (documento ou sentença) fazer parte de uma determinada classe. Esse cálculo de probabilidade é representado pela Equação 4.

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (4)$$

Pode-se calcular que um documento, *d*, tem probabilidade de fazer parte da classe *c*. $P(t_k|c)$ é a probabilidade condicional do termo *tk* do documento *d* fazer parte da classe c_1. *P(c)* é a probabilidade de o documento estar relacionado com a classe c_1.

Segundo Rish (2001), essa técnica, apesar de simples, demonstra-se eficaz em muitas aplicações práticas como, por exemplo, na classificação de texto, em diagnósticos médicos e em sistemas de gestão de desempenho.

No contexto da análise de sentimento, a aplicação do NB pode trazer grandes contribuições, como, por exemplo, o trabalho proposto por Kang, Yoo e Han (2012), que utiliza essa técnica para a adaptação de um léxico de sentimento já existente para um novo domínio.

Além do uso do NB para a adaptação de domínio, sua aplicação é muito utilizada para a própria polarização de sentenças. O motivo para o seu vasto uso é a alta taxa de acerto, como é possível observar nos trabalhos de Mihalcea e Strapparava (2006); Tan e Zhang (2008); Boiy e Moens (2009); Lane, Clarke e Hender (2012); entre outros.

Na próxima seção, demonstra-se o método *Pointwise Mutual Information* (PMI).

2.4.3.5 Pointwise Mutual Information

O *Pointwise Mutual Information* (PMI) é baseado em coocorrência e utiliza um cálculo para verificar a relação entre termos de um documento (TURNEY, 2001). Segundo Tsytsarau e Palpanas (2012), o PMI é um método de base em critérios estatísticos que trocam a probabilidade com a frequência que o termo ocorre, *F(x)*, e com a coocorrência, *F(x perto y)*, representado pela Equação 5.

$$PMI(x, y) = log_2 \frac{F(x\ perto\ y)}{F(x)\ F(y)} \quad (5)$$

O PMI não é utilizado isoladamente para a análise de sentimento e para classificação de texto. O método deve ser combinado com outras técnicas ou algoritmos. Segundo Wang, Xu e Wan (2013), o método PMI é também utilizando no contexto da análise de sentimento para extração de *collocation*, termos compostos que podem representar características de um produto ou serviço a ser avaliado. Esse recurso

também é utilizado e apresentando nos trabalhos de Liu et al. (2013), Zhang, Xu e Wan (2012).

Na visão de Cruz et al. (2013), o uso de PMI também é válido para identificar a proximidade semântica entre recursos e características de um domínio com os adjetivos polarizados de uma sentença.

Segundo Tsytsarau e Palpanas (2012), o PMI pode ser utilizado em conjunto com um léxico de opinião, a fim de encontrar a relação semântica entre os termos presentes numa sentença e os termos do léxico, previamente classificados como positivos ou negativos. Ao fim, é apresentada a probabilidade de a sentença ter determinada polarização. Para Wang e Guo (2012), o uso de PMI pode auxiliar na expansão de um léxico de sentimento, utilizando como base a correlação entre os termos para se encontrar sinônimos e auxiliar diretamente na polarização da sentença.

Alguns autores como, por exemplo, Chen et al. (2012), utilizam o PMI diretamente para calcular a orientação semântica (polarização) de termos e da sentença como um todo. Mais detalhes sobre esse e outros trabalhos podem ser vistos na Seção 2.4.5.2.

Na seção seguinte, apresenta-se a técnica conhecida como reconhecimento de entidades nomeadas.

2.4.3.6 Reconhecimento de entidades nomeadas

O reconhecimento de entidades nomeadas (ou no inglês, *Named Entity Recognition* – NER) é uma técnica de extração de informação que tem como objetivo identificar termos em textos não-estruturados, apresentando uma possível classe relacionada (ZHU; GONÇALVES; UREN, 2005).

A técnica de extração de entidades pode ser vista como um problema de classificação em que as palavras são marcadas para uma ou mais classes semânticas. Quando a entidade encontrada não pode ser assinada para uma classe específica, ela é atribuída a uma ordem ‘geral’ (KONCHADY, 2006).

O reconhecimento de entidades deve identificar as fronteiras de um termo composto, como, por exemplo, a identificação da expressão Universidade Federal de Santa Catarina como uma organização. O processo de NER pode ser executado com base em uma lista de palavras previamente separadas por classes, chamadas de *gazetteers*, conjuntamente com a confecção de algumas regras para auxiliar na desambiguação das entidades candidatas, conforme apresentado no trabalho de Ceci et al. (2010).

Caso não se tenha uma base inicial de termos, pode-se utilizar uma abordagem fundamentada em clusterização para a identificação das expressões (simples ou compostas) e submeter essas entidades candidatas à validação de um especialista (CECI; PIETROBON; GONÇALVES, 2012).

No contexto da análise de sentimento, pode-se encontrar alguns trabalhos que utilizam NER como técnica-meio ou fim para a polarização, como é possível observar nos trabalhos de Xu et al. (2011), Chen, Chen e Wu (2012), Fernandez e Losada (2012), entre outros.

Tanto o NER como as outras cinco técnicas apresentadas anteriormente representaram as principais estratégias utilizadas para a análise de sentimento e classificação semântica a partir dos resultados mostrados pela revisão sistemática.

2.4.3.7 Deep learning

Algorítmos *deep learning* (aprendizagem profunda) têm como objetivos explorar estruturas desconhecidas como entrada, a fim de descobrir boas representações, muitas vezes, em vários níveis, de modo a modelar abstrações de níveis mais altos usando arquiteturas compostas por múltiplas transformações não lineares (BENGIO, 2012).

Pode-se entender o *deep learning* como a aplicação de técnicas na área de aprendizagem de máquina (*machine learning*) para a aprendizagem de novos padrões. Segundo Arel, Rose e Karnowski (2010), os principais algoritmos para *deep learning* são passíveis de treinamento e devem possuir dados iniciais para aprendizagem.

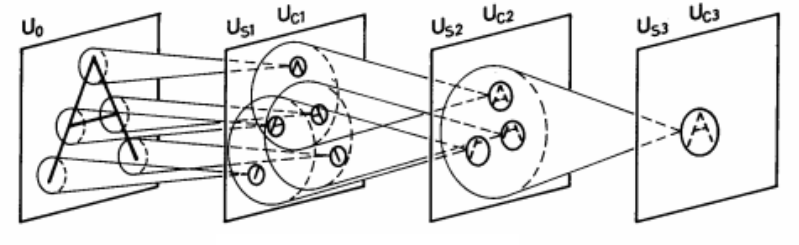
Os principais algoritmos utilizados são baseados em redes neurais artificias. Segundo Haykin (2001), uma rede neural funciona como um processador macivo, paralelamente distribuído, constituído de unidades de processamento simples que têm como possibilidade o armazenamento de conhecimento experimental e podem torná-lo disponível para o uso.

Um tipo de rede neural muito utilizado em *deep learning* é o *Convolutional Neural Networks* (CNN) ou, em português, Redes Convolucionais. Para Arel, Rose e Karnowski (2010), as CNNs fazem parte da família de redes neurais multicamadas inicialmente desenvolvidas para a utilização de dados bidimensionais, tais como imagens e vídeos. Ainda segundo os autores, essa é a primeira abordagem de *deep learning* que obteve verdadeiramente sucesso.

Na visão de Bianchini (2001), as redes convolucionais buscam atacar o problema clássico, que é falta de habilidade para se lidar com

situações em que existam deformidades nos dados de entrada. A Figura 15 apresenta um exemplo de funcionamento dessa rede.

Figura 15 – Exemplo de funcionamento das redes convolucionais



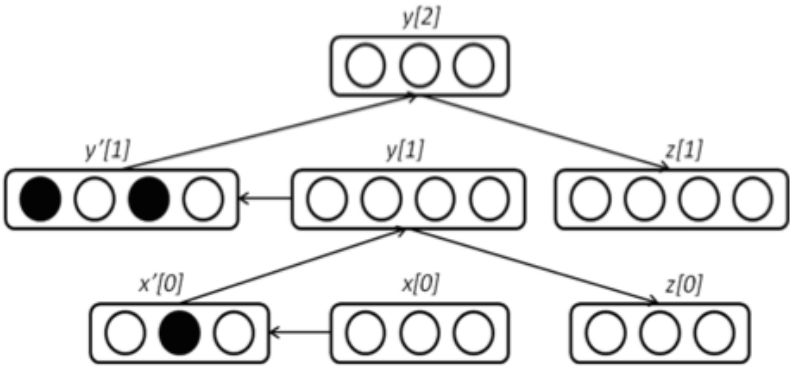
Fonte: Bianchini (2001)

Pode-se perceber, a partir da Figura 15, que as redes convolucionais trabalham quebrando e reduzindo o objeto passível de classificação em elementos menores, mantendo as características. Dessa forma, para Bianchini (2001, p. 39), "assim, múltiplos planos são usualmente utilizados em cada camada para que múltiplas características possam ser detectadas".

O *deep learning* já foi usado para a prática da análise de sentimento, como se pode verificar no trabalho de Glorot, Bordes e Bengio (2011), que procura utilizar algoritmos dessa natureza para fazer com que um conjunto treinado de classificadores em um domínio seja adaptado para um novo domínio (desconhecido).

No trabalho apresentado por Bengio (2011), é utilizado o algoritmo *Staked Denoising Autoencoders* (SDA). As redes neurais do tipo SDA têm sido utilizadas com sucesso em muitos cenários de aprendizagem de domínio. Elas atuam usando eventos randômicos para tratar os possíveis ruídos ou imperfeições nos dados de entrada da rede. A Figura 16 apresenta mais detalhes sobre o funcionamento do SDA.

Figura 16 – Exemplo de funcionamento das redes convolucionais



Fonte: You e Zhang (2013)

You e Zhang (2013) exemplificam a Figura 16 de modo que a entrada x é mapeada para y reconstruindo z. A saída y é usada como entrada para a camada de cima e o mesmo processo de iteração é executado até que se forme uma rede de profundidade, chegando à classificação desejada.

Na próxima seção são apresentados mais detalhes sobre a técnica chamada de *Singular Value Decomposition* (SDV).

2.4.3.8 Singular Value Decomposition (SVD)

Em álgebra linear, *Singular Value Decomposition* (SVD) ou, em português, decomposição de valores singulares, é um método para redução de dimensionalidade a partir de coocorrência de termos, sendo que esses termos são distribuídos em uma matriz pelas dimensões em questão (MANNING e SCHUTZE, 1999).

Ainda na visão de Manning e Schutze (1999), o processo de redução de dimensionalidade é uma técnica que pega um conjunto de objetos que existem em um espaço alto-dimensional e os modifica para a perspectiva de um espaço baixo-dimensional, muitas vezes em duas ou três dimensões para visualização.

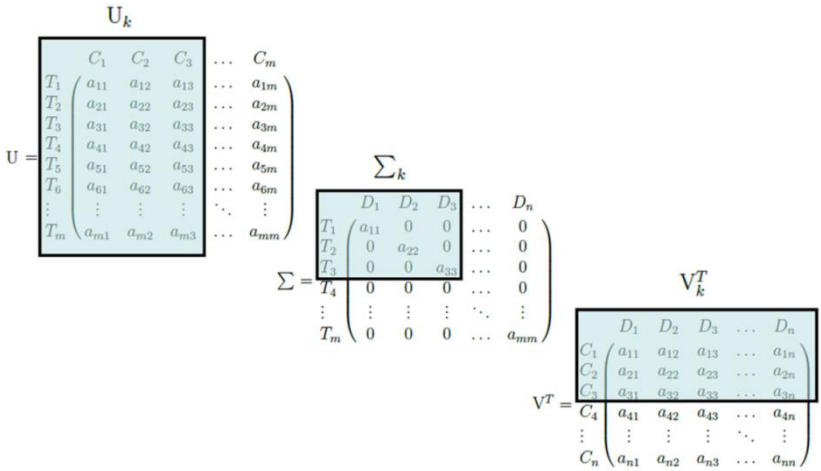
Segundo Woszezenki (2014), o SVD pode ser entendido como a fatoração de matrizes complexas. Contruindo, dessa forma, um espaço semântico em que as entidades de um domínio, fortemente relacionadas, aproximam-se. Para Manning e Schutze (1999), no processo de redução de dimensionalidade, termos que coocorrem são mapeados para

dimensões de um espaço reduzido. Sobre o funcionamento do SVD, Gonçalves (2006, p.41) explica.

> Considerando-se uma matriz esparsa termo–documento, a decomposição é calculada visando produzir uma matriz completa. Uma característica importante do modelo é a capacidade de redução de dimensionalidade através da utilização dos k fatores (valores singulares) mais relevantes. Esses fatores permitem recuperar parcialmente a informação que representa a matriz original.

Segundo Woszezenki (2014), o SVD decompõe uma matriz original M em três matrizes. A Figura 17 apresenta mais detalhes sobre esse processo.

Figura 17 – Redução da dimensionalidade da matriz original M



Fonte: Chen et al., 2013

Para entender o contexto de uso do SVD, utiliza-se a exemplificação de Tan, Steinbach, Kumar (2009). Sabe-se que uma matriz A*m* por *n* pode ser descrita conforme a Equação 6, onde σ_i é o valor singular de índice i da matriz A e u_i é o vetor singular da matriz A pela esquerda de índice i e v_i é o vetor singular pela direita do índice i.

$$A = \sum_{i=1}^{rank\,(A)} \sigma_i\, u_i\, v_i^T = U \Sigma\, V^T \tag{6}$$

Na próxima seção, apresenta-se o cenário de aplicação para a classificação semântica, mais propriamente para a análise de sentimento.

2.4.4 Contexto de aplicação

Pautados na revisão sistemática apresentada na Seção 2.1, pode-se levantar alguns contextos de aplicação para a classificação semântica, mais precisamente, da análise de sentimento.

Na seção a seguir, são apresentados contextos de aplicação para a análise de sentimento, entre eles: a análise de dados financeiros, o uso em campanhas eleitorais, a análise de *reviews*, a detecção de crime e de terrorismo e, finalmente, o *marketing*.

2.4.4.1 Análise de dados financeiros

Segundo Schumaker et al. (2012), a análise de sentimento pode auxiliar na avaliação de dados financeiros. Na visão dos autores, em notícias financeiras classificadas como negativas, fica muito mais fácil prever uma alteração nos preços do produto relacionado.

Essa lógica pode ser utilizada para previsão do mercado, ou seja, podem-se examinar notícias sobre organizações e produtos e esperar um reflexo direto em preços ou ações da bolsa de valores.

Um dos grandes desafios da análise de textos financeiros é o processamento da grande quantidade de informação que é gerada. O uso de instrumentação computacional faz-se necessário juntamente com técnicas eficientes. A análise de sentimento é uma das técnicas que pode auxiliar na avaliação dos dados, possibilitando melhores previsões (SCHUMAKER, 2012).

Na seção seguinte, é apresentado o cenário de uso de análise de sentimento para auxiliar nas campanhas eleitorais e em contextos políticos.

2.4.4.2 Uso na política

Uma área em que se tem utilizado com muita frequência a análise de sentimento é a política. Políticos e pessoas públicas em geral têm a necessidade de identificar como está a sua imagem perante o público (eleitores). A aplicação da análise de sentimento, utilizando como base textos publicados em redes sociais e no Twitter, pode trazer grandes benefícios e agilidade nessa tarefa (CHEN; CHEN; WU, 2012).

No trabalho de Park et al. (2011), é utilizada a rede social sul-coreana *Cyword*[8], na qual os políticos possuem perfis públicos, permitindo trocas de mensagens com os seus eleitores. Os autores

[8]Acesse a rede Cyword em: http://www.cyworld.com/cymain/?f=cymain

aplicaram a análise de sentimento para identificar como os eleitores veem a imagem do político, possibilitando assim ter um indicador (positivo/negativo).

No trabalho desenvolvido por Efron (2006), a mineração de opinião é utilizada de maneira diferente. Além de polarizar as sentenças, os textos são pré-classificados como de esquerda ou de direita em relação ao governo atual. Muito similar a esse trabalho, cita-se Malouf e Mullen (2008), só que, nesse caso, é possível adicionar outros classificadores que não apenas esquerda/direita, bastando fazer a definição previamente.

Alguns trabalhos relatam a utilização da análise de sentimento como método para previsão do resultado de uma eleição, conforme é possível verificar na obra de Tumasjan et al. (2011).

Chen, Chen e Wu (2012) sugerem o uso da técnica de análise de sentimento para monitorar a imagem dos candidatos e políticos já eleitos, de modo que a polarização possa ser estendida para outros qualificadores.

Na próxima seção, encontra-se exposto outro contexto de aplicação da análise de sentimentos, a análise de *reviews* ou opiniões sobre produtos e serviços.

2.4.4.3 Análise de *reviews*

O uso de análise de sentimento em avaliação de *reviews* de produtos e em *sites* de opiniões é muito comum. A seguir, são apresentadas três situações de aplicação ligadas ao contexto de *reviews* de filmes e *reviews* de hotéis.

Segundo Abbasi, Chen e Salem (2008), a *Web* é um importante canal para a disseminação de opiniões sobre produtos e serviços. Os autores propõem a utilização da análise de sentimento para classificar a imagem de um filme a partir do que as pessoas, que já o assistiram, colocam em fóruns especializados.

O trabalho apresentado por Thet, Na e Khoo (2010) utiliza a análise de sentimento não só para identificar a polarização do filme como um todo, mas também de seus elementos – diretor, trilha sonora, ator principal etc. –, além de agregar a qualificação de um grau de polarização.

Wu et al. (2010) apresenta a análise de sentimento na identificação de como está a imagem de hotéis com fundamentação nas opiniões publicadas por seus hóspedes. Na visão do autor, o resultado vindo da análise de sentimento deve ser publicado utilizando uma

estrutura gráfica que facilite a visualização das opiniões distribuídas pelo eixo tempo.

Na seção a seguir, a análise de sentimento é apresentada como ferramenta para a examinação de publicações científicas.

2.4.4.4 Análise de trabalhos científicos

A análise de sentimento pode ser utilizada para identificar o que a comunidade científica pensa de uma determinada pesquisa ou artigo, auxiliando na identificação do fator de impacto da publicação (HUANG; QIU, 2010). Os autores utilizam como base, o cruzamento das referências e, principalmente, as citações efetuadas, de modo a identificar se existe uma opinião na sentença ou não, e qual a orientação semântica dessa opinião (positiva ou negativa).

Outra aplicação da análise de sentimento nesse contexto é a proposta de Small (2011), na qual o foco é a interpretação de mapas da ciência, utilizando como embasamento as citações de artigos e identificando as coautorias. Nesse trabalho, o sentimento analisado vai além do positivo e negativo, uma vez que se procura identificar temas que são interessantes, promissores, entre outros qualificadores.

O próximo segmento apresenta o uso de análise de sentimento como ferramenta para a detecção de crimes e de ameaças terroristas.

2.4.4.5 Detecção de crimes e de terrorismo

O uso de análise de sentimento a partir dos dados disponíveis em plataformas da *Web 2.0* não é novidade. Alguns pesquisadores vêm recorrendo a esses dados para a identificação de ações de terrorismo, como se pode observar no trabalho de Yang e Dorbin (2011), no qual as informações são clusterizadas e correlacionadas de modo a se gerar um grafo.

No trabalho apresentado por Cheong e Lee (2011), a fonte principal de dados é o Twitter, e o foco de aplicação da análise de sentimento é a detecção de mensagens contendo informações sobre terrorismo. Para isso, são identificadas palavras que possam ter contexto no terrorismo a partir da classificação de uma sentença como negativa. O objetivo é prever ataques e atentados, de forma que seja possível prevenir tais eventos.

Na seção a seguir, são apresentados casos de aplicação da análise de sentimento no contexto do *marketing* e da inteligência competitiva.

2.4.4.6 Marketing

A área de publicidade e propaganda é uma das grandes beneficiadas pela aplicação da análise de sentimento. A possibilidade de identificar a opinião de usuários ou clientes sobre um produto, um serviço ou sobre a imagem da marca é algo de grande interesse nessa área.

O trabalho de Fan e Chang (2010) sugere a aplicação da análise de sentimento, com alicerce nos dados de *blogs* e do Twitter, para identificar sentimentos com o intuito de definir como está a imagem de produtos ou de serviços perante a um conjunto de usuários ou consumidores.

Xu et al. (2011) utilizaram-se da análise de sentimento para comparar opiniões a fim de auxiliar na inteligência competitiva que, segundo os autores, é um dos fatores-chave para o gerenciamento de riscos e para o apoio à decisão organizacional.

Na próxima seção são apresentados alguns métodos para a polarização de sentenças a fim de identificar o grau de polarização final.

2.4.5 Métodos para orientação semântica (polarização)

Os métodos para orientação semântica têm como objetivo apresentar o grau de polaridade do sentimento envolvido na sentença em questão. As próximas seções apresentam alguns métodos para chegar à orientação semântica de uma sentença. Na visão deste trabalho, adota-se como orientação semântica o fato de uma sentença ser positiva ou negativa, bem como o grau da orientação, que é o valor vinculado à sentença.

2.4.5.1 Combinação linear

Segundo Mejova (2011), um dos métodos mais simples utilizado é a combinação linear de todas as polaridades encontradas nos termos da sentença, como expressado nas Equações 7 e 8.

$$polarização\ (s_i) = \begin{cases} positivo, & eval(s_i) > 0 \\ negativo, & eval(s_i) < 0 \end{cases} \quad (7)$$

$$eval(s_i) = \sum_j score\ (t_j) \quad (8)$$

Nesse método, a *polarização*, ou a orientação semântica, de uma sentença, s_i, pode ser definida como positiva, caso o resultado da função *eval*(s_i) seja superior a 0 (zero), e negativa se o resultado for inferior a 0 (zero). A função *eval*(s_i) é aplicada na sentença. Para cada termo, é

somado o seu grau de polaridade, muito comumente utilizado 1 para termo t_j com propriedade positiva e -1 para termos com propriedades negativas. Ao final, é identificado o grau de polarização da sentença como resultado desse somatório.

Muitos trabalhos utilizam essa abordagem, tais como: Dave et al. (2003), Turney (2003), Mcdonald et al. (2007), Voohees e Buckland (2007), entre outros. O presente trabalho também fará uso desse método para a definição da orientação semântica (polarização).

2.4.5.2 Método Turney e Littman (2003)

O método proposto por Turney e Littman (2003) utiliza como base dois conjuntos básicos de termos positivos e negativos a fim de identificar a associação semântica entre os demais termos de uma sentença. Leva em consideração a coocorrência, aplicando o cálculo do PMI, representado pela Equação 9.

$$O(t) = \sum_{t_i \in S_p} PMI\ (t, t_i) - \sum_{t_i \in S_n} PMI\ (t, t_i) \tag{9}$$

Nessa equação, *O(t)* representa a orientação semântica de um determinado termo, levando em consideração o somatório da associação semântica desse termo com todos os demais termos presentes na sentença e no conjunto positivo, menos o somatório do conjunto negativo de termos.

Esse cálculo pode ser somado à combinação linear apresentada na Seção 2.4.5.1 a fim de atingir uma orientação semântica mais eficiente. A próxima seção apresenta um método que utiliza como base o *WordNet*.

2.4.5.3 Método proposto por Kamps et al. (2004)

O método proposto por Kamps et al. (2004) constrói um grafo baseado nas relações léxicas extraídas da sentença a partir do *WordNet*®. Para identificar a orientação semântica, é verificado o menor caminho entre os termos juntamente com os adjetivos polarizados. Entende-se como menor caminho a distância entre os dois termos, ou seja, a quantidade de termos entre eles. A Equação 10, a seguir, apresenta mais detalhes sobre essa abordagem.

$$SO(t) = \frac{d(t, mau) - d(t, bom)}{d(mau, bom)} \tag{10}$$

O adjetivo *t* é considerado positivo se *SO(t)* > 0, caso contrário ele é negativo. É efetuada uma normalização de modo que o valor da orientação semântica sempre esteja entre [-1, 1]. Esse processo é útil para identificar se os termos da sentença estão mais próximos dos termos negativos ou dos positivos baseando-se na distância das palavras no grafo gerado a partir do *WordNet*®.

2.5 ONTOLOGIA

O termo ontologia foi originalmente utilizado pelo ramo da metafísica, ciência que procura explicar a fundamental natureza das coisas, particularmente o relacionamento entre a mente e a matéria. As ontologias, na visão da filosofia, procuram estudar visões de mundo a fim de categorizar elementos. O termo foi inicialmente utilizado no ano de 1606 por Jacos Loard na sua obra *Ogdoas Scholastica* (SALM JUNIOR, 2012).

Segundo Poli e Obrst (2010), o termo ontologia hoje é visto por duas perspectivas: (1) a da filosofia, que foi mencionada anteriormente, e (2) pela perspectiva da Ciência da Computação, inicialmente utilizada pela Inteligência Artificial e hoje utilizada também pela Engenharia do Conhecimento.

Este trabalho utiliza o conceito de ontologia pela perspectiva da Ciência da Computação e da Engenharia do Conhecimento como sendo um importante artifício para a modelagem e para a representação de um conhecimento de domínio. Segundo Gruber (1993), ontologia é uma especificação explícita de uma conceitualização. O autor ainda afirma que as ontologias são uma forma para representar um conhecimento, de modo que seja possível o compartilhamento e o reaproveitamento desse conhecimento.

Para Gruber (1995), conceitualização é uma abstração, uma visão simplificada de mundo que pode ser representada. Segundo Guarino (1998), conceitualização está diretamente ligada a um domínio e a um conjunto de relações.

Na visão de Borst (1997), uma ontologia é uma representação formal e explícita de uma conceitualização comum, ou seja, compartilhada. Guarino (1998) comenta que uma ontologia é composta por um entendimento geral de um grupo para descrever certa realidade a partir de fatos conhecidos e aceitos. Para Studer, Benjamins e Fensel (1998), ontologias representam conhecimento, de forma que tanto computadores quanto humanos possam entender e raciocinar.

As ontologias definem um conjunto de representações primitivas para modelar um conhecimento de domínio (SAM; CHATWIN, 2013).

Segundo Chandrasekaran e Josephson (1999), elas representam o coração dos sistemas baseados em conhecimento, pois são responsáveis por representar o conhecimento de domínio.

As ontologias são classificadas em vários tipos, de acordo com o seu grau de generalidade ou especialidade. Guarino (1998) apresenta algumas dessas classificações:

- **Ontologias gerais** (*top-level ontology*): possuem definições abstratas para a compreensão de aspectos do mundo como, por exemplo, processos, espaços, tempo, coisas, seres etc.
- **Ontologias de tarefa** (*task ontology*): tratam de tarefas genéricas ou de atividades, como diagnosticar ou vender.
- **Ontologias de domínio** (*domain ontology*): dedicam-se a um domínio específico de uma área genérica como, por exemplo, uma ontologia sobre família.
- **Ontologias de aplicação** (*application ontology*): têm como objetivo solucionar um problema específico de um domínio, normalmente referenciando termos de uma ontologia de domínio.

Complementando a classificação apresentada por Guarino (1998), Freitas (2003) apresenta mais dois tipos:

- **Ontologias de representação**: definem as primitivas de representação, tais como *frames*, atributos, axiomas etc. na forma declarativa.
- **Ontologias centrais** (genéricas de domínio): definem os ramos de estudo de uma área ou conceitos mais abstratos dessa área.

As ontologias são compostas, de modo simplificado, por cinco elementos básicos: (1) conceitos, (2) relações, (3) funções, (4) axiomas e (5) instâncias (GRUBER, 1993; CORCHO; GOMEZ-PEREZ, 2000). As definições que se seguem foram extraídas dos trabalhos de Grube (1993), Fensel (2001) e Gomez-Perez, Fernandez-Lopez e Corcho (2004).

- Conceito ou classe: são organizados em forma de taxonomia e demostram algum tipo de interação da ontologia com a base de conhecimento.
- Relações: ilustram um tipo de interação entre as classes de um domínio.
- Funções: eventos que podem acontecer no contexto da ontologia.

- Axiomas: são verdades absolutas modeladas na forma de sentenças.
- Instâncias: representam os dados das ontologias, sendo parte de uma classe, como instância de classe.

Na visão de Studer, Benjamins e Fensel (1998), as ontologias podem ser concebidas com origem no zero ou no reaproveitando/reuso de partes ou de ontologias já existentes. O processo de criação de ontologias é algo custoso, por conta disso, algumas ferramentas computacionais foram criadas para auxiliar nessa tarefa.

Muitos autores apresentam modelos e ferramentas que visam facilitar a construção e a manutenção das ontologias de domínio a partir de conteúdo não-estruturado de determinada organização como, por exemplo: Velardi et al. (2003); Cimiano e Volker (2005); Granitzer et al. (2007); Gacitua, Sawyer e Rayson (2007); Fortuna, Lavrac e Velardi (2008); El Sayed e Hacid (2008); e Ceci, Pietrobon e Gonçalves (2012).

Na próxima seção, apresentar-se-á a importância do uso das ontologias.

2.5.1 Importância das ontologias

Segundo Beppler (2008), uma ontologia deve representar um conhecimento comum (compartilhado) de maneira consistente e não ambígua. Dessa forma, percebe-se que a ontologia se apresenta como uma forma confiável para que os sistemas baseados em conhecimento possam fazer uso da sua estrutura formal. Segundo Agarwal et al. (2015), as ontologias apresentam conceitos do domínio de aplicação, o que pode beneficiar a tarefa de identificação de entidades de destino para a análise de sentimento.

Como já foi mencionado, uma ontologia pode ser considerada o coração de um sistema baseado em conhecimento, permitindo que o domínio em questão influencie as inferências e os resultados solicitados para os seus usuários (CHANDRASEKARAN; JOSEPHSON, 1999). Freitas (2003) apresenta alguns benefícios do uso das ontologias.

- Possibilidade de reutilização do conhecimento já modelado, permitindo adaptação e extensão desse conhecimento e auxiliando na fase de construção da base de conhecimento (STUDER; BENJAMINS; FENSEL, 1998).
- Permitem consultas, comparações, verificação de consistência e integração do seu conteúdo.
- Suportam múltiplos idiomas para representar o mesmo conhecimento.

- Existe uma vasta quantidade de ontologias disponíveis para *download* e repositórios disponíveis na Internet, o que minimiza a chance de criação de uma ontologia do zero.

As ontologias podem ser aplicadas em vários seguimentos, conforme apresentado pelos autores Guarino (1998), Chandrasekaran e Josephson (1999) e Beppler (2008):

- Engenharia e representação do conhecimento;
- Modelagem de bases de dados;
- Recuperação de informação;
- Integração de dados e aplicações;
- Gestão do conhecimento;
- Bibliotecas digitais;
- Classificação de documentos e textos; entre outros.

Uma área que está se beneficiando com o uso das ontologias é a análise de sentimentos, como é possível verificar nos trabalhos de Antonini et al. (2013), Shankar e Kumar (2013), Sakthivel e Hema (2013), Borth et al. (2013), Lau, Li e Liao (2014), Penalver-Martinez (2014), Liu et al. (2015), Agarwal et al. (2015), entre outros.

Na seção seguinte, expõem-se mais detalhes sobre as ontologias utilizadas para classificação do tipo análise de sentimento.

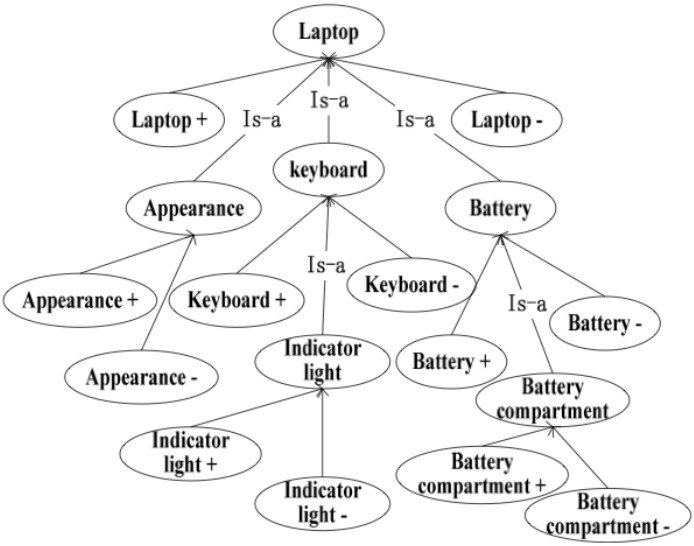
2.5.2 Ontologias de sentimentos

Segundo Sam e Chatwin (2013), uma ontologia de emoção ou sentimento define e descreve as emoções de um consumidor, ou seja, o sentimento que ele pode experimentar por um produto ou serviço.

As ontologias podem simplificar a classificação das sentenças num processo de análise de sentimento, uma vez que em suas classes existem tipos de sentimentos e emoções, trazendo atributos para facilitar a polarização.

No trabalho de Wei e Gulla (2010), apresenta-se o uso da *Sentiment Ontology Tree* (SOT). Verifica-se uma forma diferente de trabalhar com a ontologia, apresentando os produtos e os serviços organizados de maneira hierárquica e com as suas características marcadas de maneira polarizada. A Figura 18 demonstra um exemplo da ontologia SOT focada em um *laptop*.

Figura 18 – Exemplo da Ontologia SOT

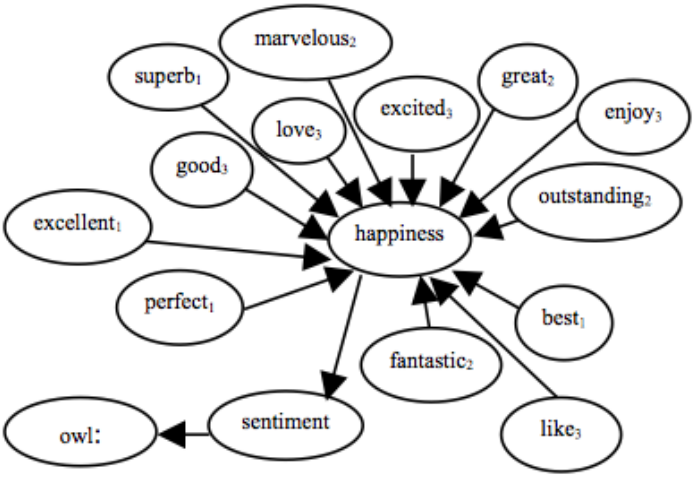


Fonte: Wang; Nie; Liu, 2013

Segundo Kontopoulos et al. (2013), no contexto da análise de sentimento, as ontologias podem ser utilizadas para mapear o domínio de aplicação bem como para demostrar o nível de emoção ou de sentimento de um adjetivo, apresentando um valor numérico para representar o seu grau. Baldoni (2012) afirma que uma ontologia de emoção ou sentimento demostra sensações estruturadas na forma de uma taxonomia e pode trazer muito mais valor para a análise do que uma simples polarização.

No trabalho de Sam e Chatwin (2013), são apresentados dois conceitos relacionados à classe sentimento: felicidade e tristeza. Para cada um desses conceitos existem vários outros que representam esse domínio de aplicação. A Figura 19 expõe os conceitos relacionados à felicidade.

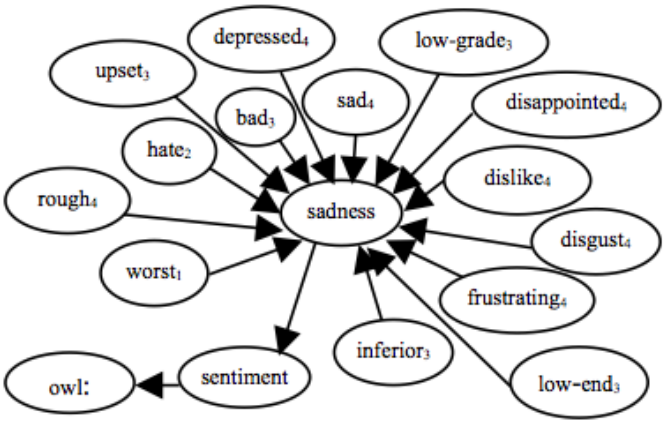
Figura 19 – Conceitos ligados à felicidade



Fonte: Sam e Chatwin (2013)

Para o conceito tristeza, é efetuado o mesmo processo, o qual está ilustrado na Figura 20.

Figura 20 – Conceitos ligados à tristeza



Fonte: Sam e Chatwin (2013)

Uma ontologia foi desenvolvida especificamente para a utilização em análise de sentimento e em mineração de opinião – a *EmotiNET*. Essa ontologia apresenta os conceitos, as relações e as propriedades

relacionados aos recursos que podem existir nos textos (BALAHUR et al., 2011).

As ontologias podem trazer muitos benefícios para a tarefa de classificação e de análise de sentimento, como, por exemplo, a representação de conceitos do domínio de aplicação, sendo que esses conceitos podem ser combinados com recursos de sentenças para auxiliar na identificação da polarização final ou no reconhecimento da entidade de destino (LIU et al., 2015). Além da possibilidade da modelagem do conhecimento de domínio, as ontologias são estruturadas de modo a permitir inferências e raciocínios a partir da sua representação.

A próxima seção apresenta mais informações a respeito das inferências e raciocínios sobre as ontologias.

2.5.3 Inferência e raciocínio

Raciocínio, na visão de Brachman e Levesque (2004), é uma manipulação formal de símbolos, que representam uma coleção de representações verdadeiras para produzir novas representações de conhecimento. Segundo Fialho (2011), normalmente o raciocínio envolve o processo de concluir e avaliar, baseado em um cálculo de argumentos.

Sobre o porquê de se trabalhar com raciocínio, Brachman e Levesque (2004) afirmam que esse recurso permite, aos sistemas baseados em conhecimento, apresentar resultados levando em consideração elementos do domínio da aplicação, além da possibilidade de se chegar a novas constatações.

Sobre inferência e raciocínio Pearl (1988) *apud* Beppler (2008) p. 34, afirmam que:

> Inferência e raciocínio normalmente são conceitos que se confundem. Raciocínio pode ser visto como o processo de inferir um novo conhecimento enquanto que inferência é a derivação em si de novos dados, fatos ou conhecimentos, que podem ser tanto positivos quanto negativos, a partir de um conjunto de dados.

Segundo Qi e Harth (2012), o uso da inferência pode auxiliar na validação da estrutura conceitual das ontologias, podendo indicar erros de modelagem e até inconsistências entre ontologias combinadas numa estratégia em rede. Para Schkegek e Shapiro (2014), é possível efetuar inferências e raciocínios a partir de grafos. Na visão dos autores, o percurso tomado dentro do grafo (ou árvore) pode caracterizar-se como

um novo raciocínio ou inferência, permitindo chegar a uma nova conclusão.

É possível visualizar as ontologias como grafos em que os conceitos e as instâncias são os nós e as relações entre as classes são as arestas. Segundo Beppler (2008), além do uso dos conceitos e das instâncias das ontologias para a inferência, pode-se utilizar regras para chegar a conclusões mais complexas.

Na seção a seguir, apresenta-se o Raciocínio Baseado em Caso, que é uma técnica focada na retenção, recuperação e reaproveitamento de casos passados para novas situações.

2.6 RACIOCÍNIO BASEADO EM CASOS

O Raciocínio Baseado em Casos (RBC) consiste na comparação de um novo problema com um já previamente resolvido para desenhar inferências que auxiliem na tomada de decisão e na resolução do dilema (WEBER; ASHLEY; BRÜNINGHAUS, 2006).

Essa técnica fornece uma metodologia para os sistemas de apoio à decisão, mais especificamente, na resolução de problemas novos com base na solução de adversidades anteriores que lhe são semelhantes. O ponto central do RBC é disponibilizar uma capacidade poderosa de aprendizagem, que busca usar experiências passadas como fundamentação ao lidar com novos problemas. Um sistema de RBC pode facilitar o processo de aquisição de conhecimento, eliminando o tempo necessário para extrair soluções dos especialistas. Segundo Niu, Lu e Zhang (2009), o método RBC parece promissor ao ser aplicado em situações dinâmicas ou nas quais se conhece pouco sobre o domínio ou sobre as soluções.

O RBC é uma técnica da Inteligência Artificial (IA) que tem como objetivo emular o conhecimento para resolver problemas seletos. Essa metodologia de resolução de problemas se diferencia de outras técnicas da IA, pois não contém apenas um conhecimento geral do domínio do problema. O RBC utiliza o conhecimento específico de uma experiência passada para resolver um problema atual. Para a resolução desse problema, é utilizada a busca de casos passados. Por meio da similaridade, podem-se identificar problemas parecidos e consequentemente soluções para o problema. Toda alteração de uma solução proposta para um problema é armazenada para posteriormente ser utilizada em novos casos (BEPPLER, 2002).

Segundo Kaster, Medeiros e Rocha (2000), as ideias do RBC foram desenvolvidas para suprir algumas desvantagens das técnicas de IA tradicionais, como, por exemplo, o Raciocínio Baseado em Regras,

que tem como função analisar uma situação a partir de regras pré-estabelecidas. Abaixo, têm-se algumas desvantagens dos sistemas baseados em regras que o RBC vem suprir, segundo o mesmo autor.

- **Obtenção de conhecimento**: é mais fácil basear-se em soluções de casos antigos e já conhecidos do que formatar uma série de regras para expressar um conhecimento.
- **Memória**: sistemas baseados em regras não possuem memória, ou seja, sempre iniciam a análise de um problema do zero. Já nos sistemas baseados em casos, busca-se um problema anteriormente resolvido, similar ao atual, a fim de auxiliar a solução do problema atual.
- **Robustez**: se o sistema baseado em regras não tiver normas compatíveis com o problema, ele simplesmente deixa de resolvê-lo. Nos sistemas baseados em casos, uma solução conhecida pode ser adaptada para atender a resolução de um novo problema.

Na visão de Gunawardena e Weber (2012), o RBC é utilizado em vários campos, nos quais a tarefa foco seja raciocínio, diagnóstico, classificação e recomendação.

2.6.1 Etapas de um RBC

Segundo Aamodt e Plaza (1994), são quatro as principais etapas de um RBC: (1) recuperar, (2) reutilizar, (3) revisar e (4) reter. A seguir, são descritas as características de cada uma delas.

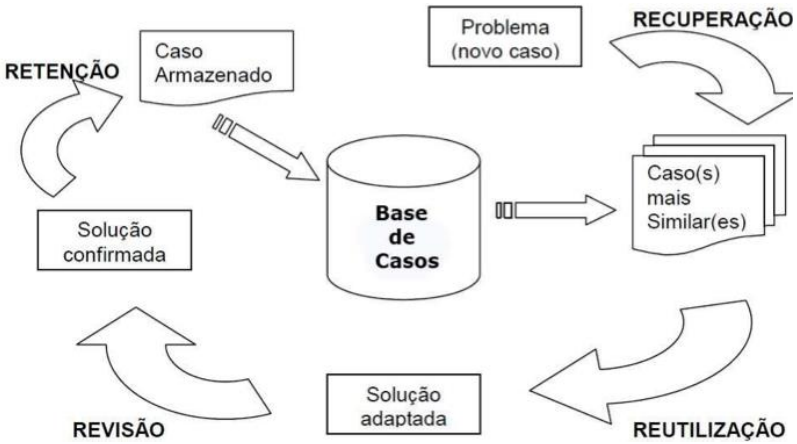
- **Recuperar**: o objetivo dessa etapa é recuperar os casos que tenham uma solução mais semelhante ao problema apresentado. Ela tem início na descrição de um problema e termina quando o caso com a maior similaridade é encontrado (AAMODT; PLAZA, 1994; BEPPLER, 2002).
- **Reutilizar**: a partir do caso recuperado, a reutilização foca em dois aspectos – as diferenças entre os casos antigos e os atuais e qual parte do caso recuperado pode ser aproveitada para a nova circunstância (AAMODT; PLAZA, 1994). Segundo Beppler (2002), os casos recuperados precisam de certa modificação para atender aos novos problemas, essa situação é chamada de adaptação.
- **Revisar**: nessa etapa, verifica-se se a reutilização está correta e, caso existam problemas, aprende-se com as falhas. Ela é dividida em duas etapas: avaliação da solução gerada para o

reuso e reparação da solução usando o conhecimento específico do domínio (BEPPLER, 2002; AAMODT; PLAZA, 1994).

- **Reter**: quando o caso é resolvido e o processo de revisão efetuado, esse caso pode ser armazenado. O sistema de RBC só é eficiente se aprender a partir das experiências passadas e da correta indexação dos problemas (BEPPLER, 2002).

A Figura 21 apresenta como as etapas aqui descritas constituem o chamado ciclo do RBC.

Figura 21 – Ciclo do RBC



Fonte: Adaptado de Aamodt e Plaza (1994)

Segundo Ji et al. (2013), o ciclo do RBC é composto pelas suas principais etapas: recuperar, reutilizar, revisar e reter. Ademais, a técnica é conhecida como ‘os quatro R’s’.

2.6.2 Representação de casos

Segundo Junior et al. (2006), os casos, num sistema RBC, são responsáveis por representar o conhecimento do especialista para a resolução de um determinado problema. Os mesmos autores definem caso como sendo uma experiência real em que o problema proposto já foi resolvido, buscando representar o problema por atributos. Esses atributos têm como função destacar o contexto e o conteúdo do problema para auxiliar na sua recuperação. Para Kolodner (1993), são três o número de componentes para representar um caso:

(1) **Descrição do problema**: listar as principais características para a identificação do problema.

(2) **Descrição da solução**: solução recuperada pelo sistema para o problema em questão, a qual já deve estar validada.
(3) **Conclusão**: resultados de uma avaliação da solução durante a sua aplicação a fim de demonstrar os resultados obtidos.

Segundo VonWangenheim e VonWangenheim (2003), a representação do conhecimento de casos define os formalismos com os quais é formulado o conhecimento para um novo problema.

2.6.3 A recuperação e a indexação de casos em RBC

Tanto a recuperação quanto a indexação são fundamentais para a construção de um sistema de RBC. Mesmo sendo diferente, uma etapa precisa da outra para ter utilidade. A seguir, é apresentado um detalhamento dessas duas etapas.

A indexação consiste basicamente em criar índices dos atributos e dos termos em geral, a fim de auxiliar no processo de recuperação dos casos. Por outro lado, segundo Junior et al. (2006), a fase de recuperação tem como função encontrar o melhor caso a partir de um problema ou situação. Para isso, é necessário verificar a similaridade do caso com o problema apresentado, precisando-se varrer o índice criado na fase da indexação.

Beppler (2002) explica que, para calcular a similaridade, existem funções numéricas. Geralmente a mais usada é a "vizinho mais próximo", representada pela equação declarada a seguir, em que *T* é o caso alvo, *S* é o caso fonte, *n* é o número de atributos em cada caso, *i* é cada atributo individual, variando de 1 a *n*, *f* é a função de similaridade para o atributo *i* no caso *T* e *S*, e *W* é o peso relativo ao atributo *i*.

$$similaridade\ (T,S) = \sum_{i=1}^{n} f(Ti,Si)Wi \qquad (11)$$

O conceito de similaridade é muito importante para o RBC. Segundo a hipótese de VonWangenheim e VonWangenheim (2003, p. 96), "problemas similares possuem soluções semelhantes".

2.6.4 Uso de RBC com ontologias

Na visão apresentada por Bergmann e Schaaf (2003), o uso de raciocínio baseado em caso combinado com ontologias pode trazer um grande benefício no processo de resolução de problemas, já que é possível levar em consideração o conhecimento de domínio.

As ontologias podem ser utilizadas como forma de integração entrem bases de conhecimento e os casos de um sistema de RBC. No trabalho de Roth-Berghofer e Adrian (2010), utiliza-se as ontologias para mapear os conhecimentos disponíveis em bases ligadas na *Web* (*web linked data*) para agregar informações aos casos conhecidos da base de casos.

Para Heitmann e Hayes (2010), as ontologias podem ser utilizadas para representar um caso, de modo que este possa ser integrado com o conhecimento de domínio. No trabalho de Gaillard et al. (2013), elas são utilizadas para auxiliar na recuperação dos casos, tentando buscar a relação entre o texto do problema e os conceitos do domínio.

Percebe-se que o uso de conceitos do domínio, em estratégia de RBC, é algo trabalhado por muitos autores. Na seção sucessora a esta, são analisadas soluções que combinam o uso de RBC com a análise de sentimento.

2.6.5 RBC para auxiliar análise de sentimento

O uso do RBC pode ser combinado com a análise de sentimento para diferentes aplicações. No trabalho de Dong et al. (2013a), a análise de sentimento é combinada ao RBC para o desenvolvimento de um sistema de recomendação de produtos. Na visão dos autores, pode-se utilizar a supracitada análise como uma forma de extração dos pontos positivos e negativos de um produto. Essas informações são combinadas com os casos armazenados pelo RBC com o propósito de apresentar uma recomendação mais eficiente.

Pode-se utilizar RBC ainda com a análise de sentimento para gerar um *review* de funcionalidades dos produtos com base em opiniões expressadas por usuários. Nesse trabalho também utilizam-se ontologias para ter uma visão do contexto em que o produto faz parte (DONG et al., 2013b).

No trabalho de Li e Tsai (2013), utiliza-se RBC para auxiliar a análise de sentimento, ou seja, a polarização é o objeto de estudo. Os autores combinam o uso de RBC com técnicas de classificação baseada em conjutos difusos (*Fuzzy sets)* para chegar à conclusão se uma sentença ou documento é positivo ou negativo. A pesquisa realizada nesta tese fará uso de RBC com o mesmo foco do trabalho de Li e Tsai (2013). Ou seja, pretende-se entregar, ao final do processo, uma sentença ou um documento polarizado.

2.7 CONSIDERAÇÕES FINAIS

O presente trabalho tem como foco principal a classificação semântica de documentos e textos, sendo que, dentre os vários domínios de aplicação da classificação semântica, optou-se por focar na análise de sentimento e na mineração de opinião.

Desenvolveu-se uma revisão sistemática para identificar um *gap* na proposta de solução para o problema de classificação de documentos no atual estado da arte. Tem-se como objetivo reaproveitar os raciocínios passados para auxiliar nas novas classificações.

Identificou-se que o uso de RBC e de ontologias combinadas a técnicas de aprendizagem de máquina pode auxiliar na construção de um modelo para a proposta de solução deste trabalho.

Neste capítulo, além das definições de classificação semântica, sentimento e análise de sentimento, também foram apresentados conceitos relacionados com raciocínio baseado em caso e ontologias. O próximo capítulo tem como função apresentar o modelo inicial, proposto, a partir dos estudos realizados e relatados neste capítulo de referencial teórico, para abordar o problema de pesquisa.

3 MODELO INICIAL PROPOSTO

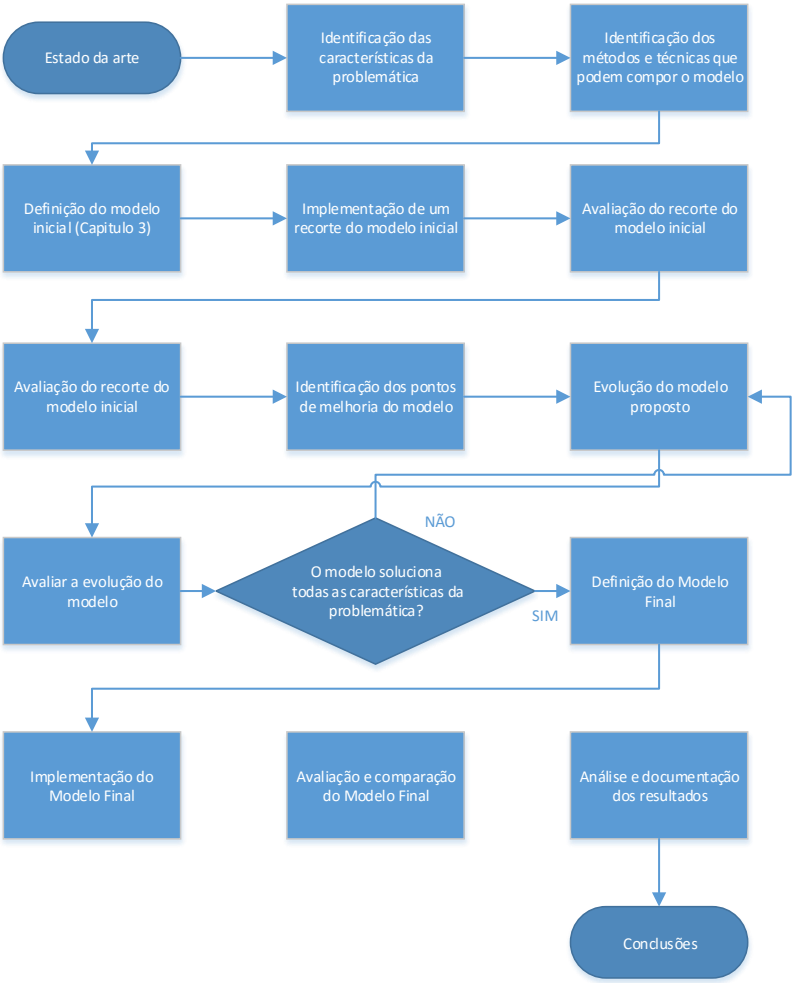
Este capítulo apresenta, de maneira detalhada, o modelo inicial proposto, o qual tem como objetivo demonstrar uma solução à problemática exposta neste trabalho bem como promover uma resposta ao *gap* de pesquisa, identificado por meio da revisão sistemática. O modelo apresentado neste capítulo não tem como objetivo ser o modelo final de tese. A função deste modelo é apresentar a inspiração inicial da proposta de solução, levando em consideração os elementos apresentados nos Capítulos 1 e 2.

O Capítulo 3, portanto, está dividido em três seções. Primeiramente há uma introdução, a qual é seguida pela descrição detalhada do modelo e das suas etapas, terminando com a apresentação das considerações finais.

3.1 INTRODUÇÃO

Para construir um modelo de tese, é necessário utilizar como base a pergunta de pesquisa e os elementos da problemática, de modo que o modelo apresente soluções e respostas. Deve-se também levar em consideração os elementos e as características descobertas no estado da arte das temáticas deste trabalho. A Figura 22 apresenta o procedimento metodológico.

Figura 22 – Procedimento metodológico para se chegar ao modelo final



Fonte: Elaborado pelo autor

Para a construção do modelo inicial, utilizou-se como base as técnicas e as ferramentas explicitadas na revisão sistemática. Dessa forma, o modelo inicial faz uso da estrutura do raciocínio baseado em casos, tendo a disposição das suas etapas organizadas de maneira similar ao RBC. Quando uma sentença ou documento é submetido ao processo de classificação, a primeira tarefa a ser executada é o reconhecimento de entidades nomeadas, visando criar subsídios para identificar o domínio de aplicação.

Para a representação e o armazenamento do conhecimento de domínio, utiliza-se uma ontologia que tem como função trazer elementos do domínio de aplicação para a classificação.

O modelo é flexível para a utilização de técnicas de aprendizagem de máquina, não tendo dependência direta com nenhuma delas. Portanto, fica a cargo do usuário escolher qual a melhor técnica para o seu problema.

O processo de POS *Tagger* é aplicado para auxiliar na estratégia de polarização escolhida, disponibilizando as classes gramaticais de cada termo de uma sentença.

Na seção a seguir, são apresentados mais detalhes sobre os requisitos funcionais que o modelo de tese deve respeitar, os quais foram explicitados na revisão teórica.

3.2 REQUISITOS FUNCIONAIS PARA O MODELO

Para a construção do modelo proposto, levantou-se os principais requisitos funcionais que devem ser atendidos a partir dos elementos identificados na revisão sistemática e na revisão bibliográfica. Esta seção tem como objetivo apresentar esses requisitos funcionais, de modo que a construção do modelo seja executada da maneira mais adequada. A seguir são apresentados os requisitos levantados a partir dos Quadro 1,Quadro 2, Quadro 3 eQuadro 4.

Quadro 1 – Requisito RF001

RF001 – O modelo deve ser sensível ao domínio.
Descrição: Para atingir um melhor resultado, a solução deve ser sensível ao domínio, ou seja, deve levar em consideração elementos do domínio de aplicação.

Fonte: Elaborado pelo autor

O requisito RF001 foi retirado de uma afirmação utilizada na problemática deste trabalho, a qual explica que a polarização de uma senteça em positiva e negativa está sujeita a uma taxa alta de erros caso não leve em consideração elementos do domínio (ZHANG; LIU, 2011a).

Quadro 2 –Requisito RF002

RF002 – O modelo deve apresentar o "porquê" de ter chego a uma polarização.
Descrição: A solução deve apresentar de maneira clara quais os motivos que levaram uma sentença a ser classificada como positiva ou negativa.

Fonte: Elaborado pelo autor

O requisito RF002, foi inspirado na problemática deste trabalho, a qual afirma que apresentar resultados da polarização apenas como positivo ou negativo não é o suficiente para a tomada de decisão. Muitas vezes o caminho até a classificação é mais importante que o próprio resultado (LI; XIA; ZHANG, 2011).

Quadro 3 –Requisito RF003

RF003 – O modelo deve 'aprender' (armazenar e recuperar) a partir de polarizações passadas.
Descrição: A solução deve estar preparada para armazenar novas polarizações efetuadas e aproveitar as classificações passadas para auxiliar na polarização de um novo conteúdo (caso).

Fonte: Elaborado pelo autor

O requisito RF003 apresenta-se a partir de um item levantado na problemática e inspirado no trabalho de Kaiser, Schlick e Bodendorf (2011). Segundo os autores, para ter uma análise de sentimento mais efetiva, é necessário 'aprender' a partir das práticas que obtiveram sucesso.

Quadro 4 –Requisito RF004

RF004 – O modelo deve permitir que sejam armazenadas as inferências realizadas com sucesso.
Descrição: A solução deve estar preparada para reutilizar as inferências elaboradas para chegar a uma polarização com sucesso.

Fonte: Elaborado pelo autor

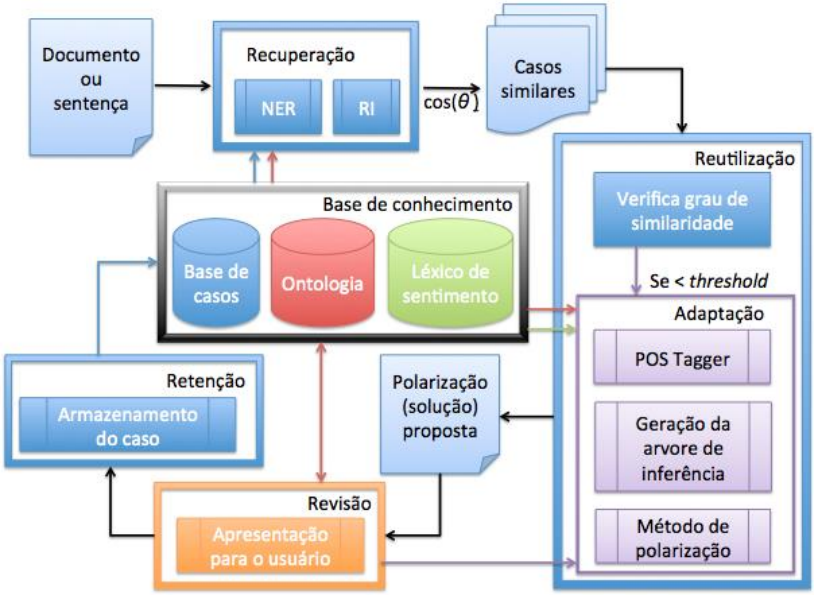
O requisito RF004 foi inspirado da afirmação efetuada por Feng et al. (2011), ao comentar que para cada nova classificação é necessário considerar todo o processo de inferência e de classificação novamente, não aproveitando todo o raciocínio já realizado.

A próxima seção utilizou como base os requisitos funcionais aqui ilustrados com o objetivo de apresentar mais detalhes sobre o modelo proposto.

3.3 DESCRIÇÃO DO MODELO

O modelo proposto tem como objetivo propor um caminho para a implementação de uma solução com o intuito de viabilizar a análise de sentimento de sentenças ou de documentos. A Figura 23 apresenta mais detalhes sobre o modelo.

Figura 23 – Proposta de solução



Fonte: Elaborado pelo autor

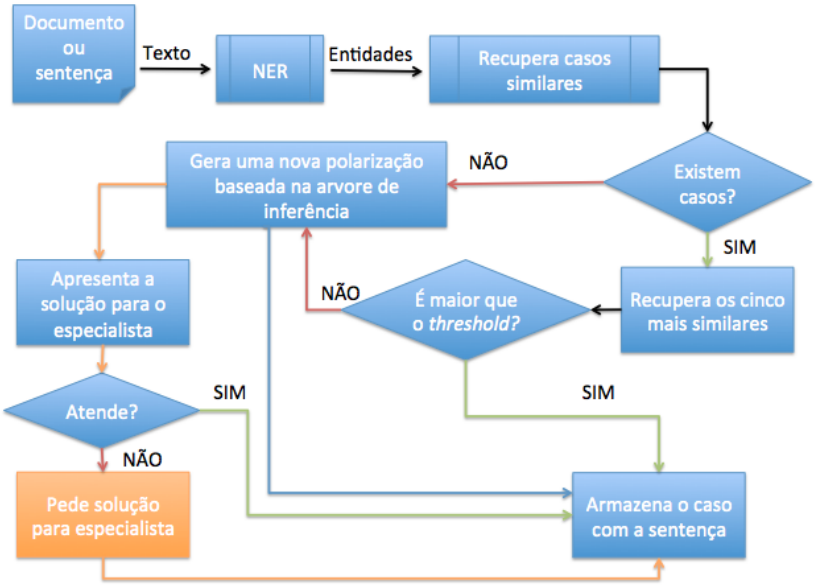
Como entrada, espera-se um documento não-estruturado ou uma sentença que pode ser de qualquer domínio. O modelo foi estruturado com fundamentação nas quatro etapas do ciclo do RBC: (1) recuperação; (2) reutilização; (3) revisão; e (4) retenção.

Cada etapa possui responsabilidades e técnicas específicas para atingir os objetivos propostos neste trabalho, o qual tem foco na classificação de sentenças e utiliza como cenário a análise de sentimento.

A inspiração para o uso do ciclo RBC como plano de fundo para o modelo proposto ocorreu por conta da habilidade do modelo em armazenar, recuperar, adaptar e reutilizar casos passados, recurso esse que vai de encontro com uma das deficiências encontradas nos modelos atuais para classificação e análise de sentimento – a falta do reaproveitamento das inferências já realizadas anteriormente.

O fluxograma, apresentado na Figura 24, traz mais detalhes de como as etapas do modelo estão encadeadas.

Figura 24 – Fluxograma da proposta de solução.



Fonte: Elaborado pelo autor

O modelo proposto permite a interação de um usuário na etapa de revisão. Essa etapa não é obrigatória para a implementação e utilização do modelo, tornando a abordagem para a classificação automática, sem a etapa de revisão, ou semiautomática, ao incluir esse passo. O fluxograma apresentado na Figura 24, demonstra o encadeamento das etapas no modelo semiautomático, ou seja, que abriga todos os quatro passos do RBC.

As próximas subseções exibem um detalhamento sobre cada uma das etapas do modelo proposto.

3.3.1 Recuperação

A partir da submissão de um documento ou sentença não-estruturada, a primeira etapa do modelo é a de recuperação. Nela são utilizadas as técnicas de reconhecimento de entidades (na Figura 23 representada como NER) e recuperação de informação (RI).

A primeira etapa consiste na aplicação de técnicas de reconhecimento de entidades, utilizando como fundamento, recursos armazenados na base de conhecimento da organização. O objetivo dessa

etapa é identificar as entidades e classificá-las (classe da ontologia) para selecionar qual é o sujeito da sentença.

A partir da identificação do sujeito e das demais instâncias e classes da ontologia, constrói-se uma *query*, ou seja, uma consulta para que a classe seja submetida ao processo de recuperação dos casos. O processo de recuperação dos casos é realizado com alicerce na área da recuperação de informação, utilizando, como base, o cálculo de similaridade do cosseno[9].

A seguir, apresenta-se o fluxo da Etapa 1 por meio de um algoritmo:

- **Passo 1**: o documento ou sentença de entrada é informado;
- **Passo 2**: o conteúdo é submetido ao processo de reconhecimento de entidades;
- **Passo 3**: as entidades reconhecidas a partir da ontologia de domínio são retornadas;
- **Passo 4**: caso não sejam encontradas entidades, siga para o **Passo 7**, caso contrário, siga para o **Passo 5**;
- **Passo 5**: são buscados casos que possuam as entidades reconhecidas (instâncias e classes da ontologia) em seu registro;
- **Passo 6**: são retornados os casos previamente armazenados;
- **Passo 7**: é concluída a etapa de recuperação.

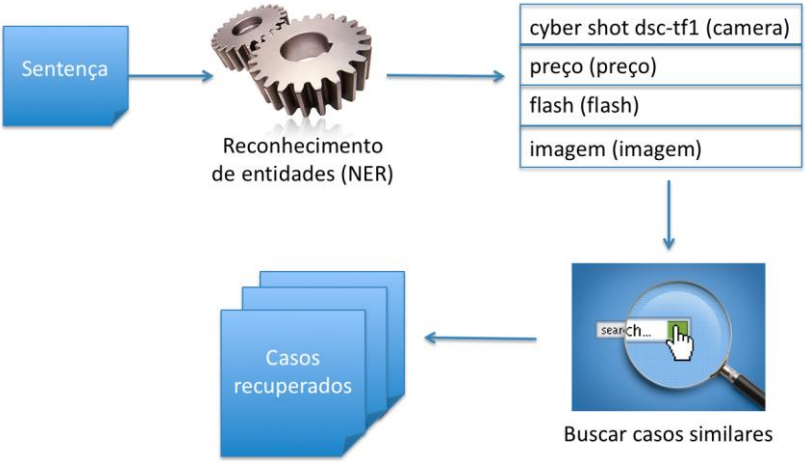
Para facilitar o entendimento dessa etapa, formulou-se o seguinte exemplo. Tem-se, como entrada, a seguinte sentença: "A *cybershot*[10]*dsc-tf1* é um ótimo equipamento, ela se apresenta ao mercado com um bom preço, seu *flash* é muito bom e a imagem tem uma ótima qualidade.".

Os próximos passos do exemplo são detalhados na Figura 25, a seguir.

[9] Este trabalho utiliza para o cálculo do cosseno a ferramenta Apache Lucene (https://lucene.apache.org/core/)

[10] Cybershot é um nome de produto registrado da marca Sony.

Figura 25 – Exemplo da etapa de recuperação do modelo proposto



Fonte: Elaborado pelo autor

A sentença submetida ao processo de reconhecimento de entidades teve como retorno quatro elementos: a instância *cybershot dsc-tf1*, que faz parte da classe câmera da ontologia de domínio, e outros três conceitos também presentes na ontologia: preço, *flash* e imagem. Tendo os quatro elementos de retorno, formula-se a consulta utilizando tais informações como, por exemplo, *(camera:cyber shot dsc-tf1) AND (flash) AND (preço) AND (imagem).*

A consulta representa uma busca por casos já existentes na base de casos que sejam relacionados à câmera *cybershot dsc-tf1* e que, no seu processo de raciocínio, tenham sido levados em consideração os conceitos (características) de *flash*, preço e imagem. Na sequência, é utilizado um processo de similaridade baseado no TF-IDF e, ao final, são retornados os casos similares com os elementos em questão.

Os casos revistos são submetidos à segunda etapa do processo, a reutilização. A próxima seção apresenta mais detalhes.

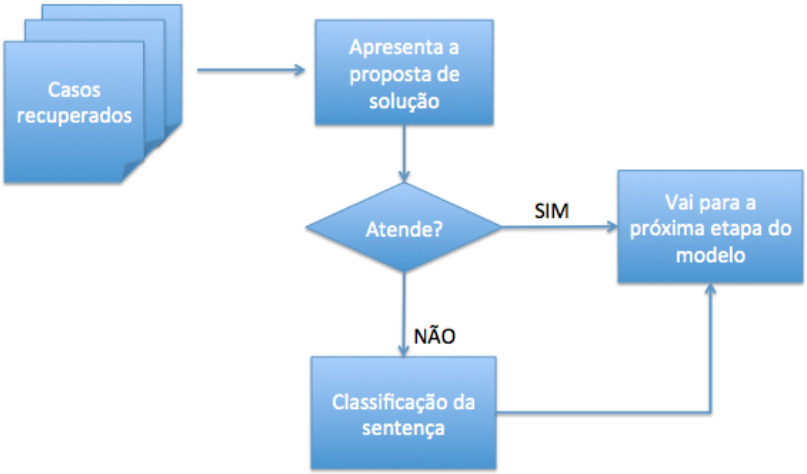
3.3.2 Reutilização

A etapa de reutilização do modelo proposto pode utilizar, como base, os casos resultantes, quando recuperados, ou fazer a classificação a partir do início.

O processo de reaproveitamento dos casos resultantes da etapa de recuperação é bastante simples. Basicamente, as ações são organizadas para a apresentação da mesma solução dos casos recuperados como solução do caso atual. Se não forem encontrados casos similares, ou se

os casos encontrados não atendam ao novo caso, a etapa de reutilização executa o processo de classificação. A Figura 26 expõe mais particularidades desse fluxo.

Figura 26 – Fluxo da etapa de reutilização



Fonte: Elaborado pelo autor

Caso não seja implementada uma estratégia semiautomática, a etapa de verificação é ignorada. Deve-se, neste caso, definir um *threshold*, de forma que, se a solução encontrada (caso recuperado) tiver um grau superior ao valor definido (*threshold*), ela é apresentada como proposta de solução.

Para o processo de classificação, inicialmente é aplicada a técnica de POS *Tagger*, que tem como objetivo identificar todas as classes gramaticais dos termos presentes nas sentenças. Após ter todos os termos com suas respectivas classes gramaticais vinculadas, o conteúdo é submetido ao processo de quebra de sentença. Esse processo tem como foco separar o texto original em pequenas sentenças, com o propósito de facilitar o processo de classificação das entidades encontradas (localização do sujeito da sentença).

Cada sentença resultante do processo de quebra de sentenças terá seus adjetivos selecionados, juntamente com a entidade (caso exista). Para saber qual adjetivo está vinculado a qual entidade, utiliza-se uma estratégia baseada na distância entre duas palavras, inferindo que o adjetivo está relacionado à entidade mais próxima. Para facilitar o entendimento, é apresentada a continuação do exemplo, iniciado na

etapa de recuperação, o qual teve como sentença de entrada o seguinte texto: “A *cybershot dsc-tf1* é um ótimo equipamento, ela se apresenta ao mercado com um bom preço, seu *flash* é muito bom e a imagem tem uma ótima qualidade.”.

Já é sabido que, nessa sentença, existem as seguintes entidades (resgatadas na etapa de recuperação): *cybershot dsc-tf1* (câmera), *flash*, preço e imagem. O processo de POS *Tagger* identifica as classes gramaticais dos termos. Submetendo a sentença original ao processo de quebra de sentença, têm-se duas sentenças resultantes.

(1) “A *cybershot dsc-tf1* é um ótimo equipamento, ela se apresenta ao mercado com um bom preço.”; e

(2) “seu flash é muito bom e a imagem tem uma ótima qualidade.”.

A partir disso, são recuperadas as entidades e os adjetivos (previamente selecionados no processo de POS *Tagger*) com suas posições em cada sub-sentença, como se pode observar a seguir.

Sentença 1: *cybershot dsc-tf1 (*POS = *2)*; *ótimo (*POS = *7)*; *bom (*POS = *16)*; e *preço (*POS = *17)*

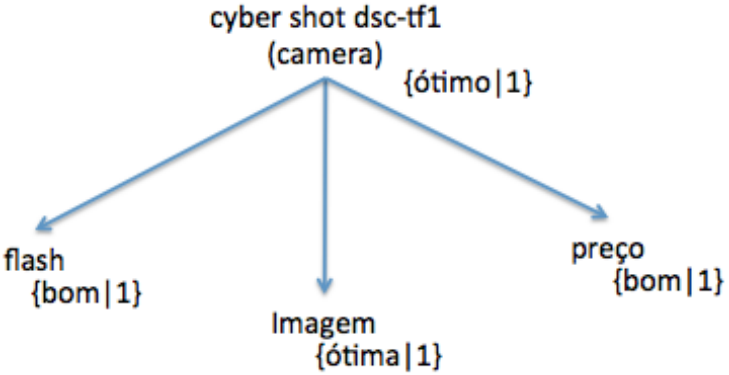
A distância entre a entidade *cybershot dsc-tf1* e o adjetivo ótimo é de 5 palavras, já a distância da mesma entidade para o adjetivo bom é de 14 palavras. Tendo em vista esse cenário, infere-se que o adjetivo ótimo está relacionado à entidade *cybershot dsc-tf1*.

O cálculo da distância pode ser representado a partir da seguinte equação: $distância\ (w1, w2) = pw2 - pw1$. Sendo que a distância do termo 1 (*w1*) para o termo 2 (*w2*) é calculada pela posição do termo 2 (*pw2*) menos a posição do termo 1 (*pw1*) na sentença.

Após identificar os adjetivos que estão vinculados aos conceitos (entidades) recuperados das sentenças, é executada uma consulta em um léxico de sentimento, em que é observado se o adjetivo tem conotação positiva (valor 1) ou negativa (valor -1).

O próximo passo consiste na recuperação do caminho entre os conceitos, por meio da navegação da ontologia de domínio. Para cada definição, deve-se somar o valor da sua polarização, a fim de chegar ao valor final do conceito-pai (ou raiz) da árvore. A Figura 27 apresenta um exemplo dessa árvore.

Figura 27 – Árvore de inferência baseada na SOT



Fonte: Elaborado pelo autor

O resultado da polarização do documento exemplificado é positivo com o valor = 4, proveniente do somatório dos valores atribuídos a cada conceito que apareceu na árvore de inferência resultante. Esse processo de representação foi inspirado nos trabalhos de Wei e Gulla (2010) e Wang, Nie e Liu (2013), nos quais eles referenciam como SOT, *sentimento ontology tree*.

Esta tese utiliza apenas a forma de representação para o seu processamento, neste caso, são aplicados conceitos de caminhamento[11] em árvores, vindos da área de estrutura de dados.

Ao final dessa etapa, tem-se uma proposta de polarização para a sentença inicialmente submetida. A próxima subseção apresenta a revisão que, se existente, caracteriza a implementação do modelo como semiautomático, pois necessita da participação de um especialista.

3.3.3 Revisão

A etapa de revisão é a única que não é obrigatória para a implementação do modelo e define se o modelo será automático ou semiautomático. Essa etapa tem como objetivo permitir a participação de um especialista do domínio no processo de classificação, momento em que é possível validar as qualificações já efetuadas, permitindo que a implementação aprenda com a interação do usuário.

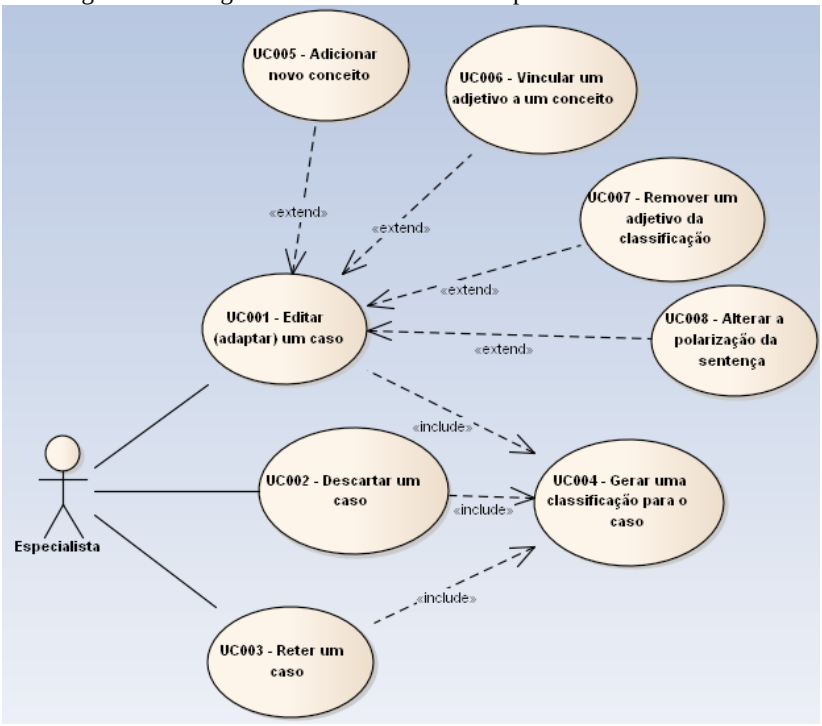
[11] Caminhamento no contexto desta tese diz respeito aos caminhos percorridos em uma árvore (grafo) para atingir um raciocínio. Também pode ser considerado como o caminho entre dois nós de um grafo.

A participação do usuário pode ocorrer diretamente com os casos, mas o reflexo dessa ação pode ter impacto direto na ontologia de domínio, ou seja, a interação do usuário nesse momento pode adicionar conceitos e instâncias à base de conhecimento, fazendo com que ela seja constantemente atualizada.

Para que a etapa da revisão seja viável, é necessário que se desenvolva uma interface gráfica para que o especialista de domínio possa interagir com os elementos dos casos disponíveis. Para facilitar a representação das possibilidades de operação dessa interface gráfica, formulou-se um diagrama de casos de uso. Segundo Bona (2002), tal diagrama tem como objetivo ilustrar o que o usuário pretende fazer com o sistema. Pode-se observar que o diagrama mostra ações que o sistema faz para suprir as necessidades que se pretende resolver.

O diagrama de casos é um dos diagramas da UML. A Linguagem de Modelagem Unificada (do inglês, *Unified Modeling Language*) consiste numa notação para modelagem de sistemas (LARMAN, 2000). A Figura 28 apresenta o diagrama de caso de uso.

Figura 28 – Diagrama de casos de uso da etapa de revisão do modelo



Fonte: Elaborado pelo autor

O especialista, a partir da interface gráfica, pode efetuar três operações básicas: (1) editar (adaptar) um caso, (2) descartar um caso ou (3) reter um caso. Contudo, para que essas operações possam ser executadas, é necessário que, anteriormente, tenha sido gerada uma classificação para o caso (representado pelo caso de uso UC004 do diagrama). Uma opção disponível para o especialista é a de reter um caso (UC003), sendo possível salvar uma categorização (polarização gerada). Essa operação pode acontecer a partir da sugestão de um processo já armazenado e recuperado como proposta de solução para o novo caso, ou ainda, por meio da edição por parte do especialista.

A operação de descarte (UC002) de uma ocorrência ocorre quando não se chega a uma classificação satisfatória e o usuário tem a possibilidade de desconsiderar o caso, ou seja, o caso não será adicionado à base de casos.

A última operação que o especialista pode desempenhar mediado pela interface gráfica é a de edição de um caso (UC001), ou seja, permite-se que o especialista valide uma classificação ou altere a categorização do evento apresentado como solução para a sentença (adaptação). Ao selecionar a operação de edição, é possível executar as seguintes operações:

- Adicionar um novo conceito (UC005): permite que o especialista identifique um novo conceito (característica) do domínio a partir do conteúdo do documento submetido para classificação. Esse processo auxilia na manutenção da ontologia de domínio.
- Vincular um adjetivo à um conceito (UC006): o entendedor pode verificar se os adjetivos encontrados estão relacionados com os conceitos identificados de maneira correta.
- Remover um adjetivo da classificação (UC007): possibilita a remoção de um adjetivo, para que o ele não influencie no cálculo da polarização.
- Alterar a polarização da sentença (UC008): permite que o especialista altere o valor final da polarização da sentença.

A revisão é seguida pela retenção, que é a última etapa do modelo. Têm-se mais informações sobre essa etapa na próxima seção.

3.3.4 Retenção

A etapa de retenção é a última do ciclo previsto no modelo da tese. É por meio desse estágio que é possível que a implementação do

modelo aprenda com os casos passados. O seu foco é formatar as informações na forma de um caso que possa ser armazenado, recuperado e estendido, de forma que essas informações sejam conservadas numa base de casos.

Para que um caso seja expressivo, de modo que também seja útil para as etapas do modelo, é necessário que ele possua os seguintes dados:

- Identificador: um número sequencial que apresente distinção entre os casos armazenados;
- Polarização/classificação: um campo que armazene a classificação (polarização) dada para as sentenças;
- Caminho percorrido na ontologia: deve armazenar qual o caminho (conceitos utilizados) percorrido na ontologia para chegar à polarização, juntamente com os adjetivos relacionados;
- Sentença original: guardar o texto original da sentença que gerou a classificação; e
- Lista de casos similares: lista de referências dos casos que possuem uma "solução" similar ao evento em questão.

Para a implementação da solução proveniente do modelo, sugere-se que, como estrutura de armazenamento dos casos, opte-se por uma construção baseada em índice invertido para facilitar a aplicação do cálculo de similaridade baseado no cosseno, proposto por esse modelo, como base para a recuperação dos casos.

A próxima seção exprime as considerações finais sobre o modelo relatado nesta tese.

3.4 CONSIDERAÇÕES FINAIS

Este capítulo teve como objetivo apresentar o modelo que visa a classificação de documentos não-estruturados para análise de sentimento. O modelo possui quatro etapas, das quais apenas uma delas é optativa para a sua implementação. Os passos do modelo foram inspirados nas etapas do ciclo clássico de RBC, que são: (1) recuperação, (2) reutilização, (3) revisão e (4) retenção.

Primeiramente, ocorre a recuperação, sendo essa etapa a responsável pelo resgate de casos já conhecidos, de modo que seja possível apresentá-los como proposta de solução para novos casos. Se não for encontrado nenhum caso relacionado, deve-se submeter a sentença original à próxima etapa.

A segunda etapa do modelo é a reutilização. Inicialmente são expostos os casos recuperados na etapa anterior. Porém, se não foram

encontrados casos, ou os recuperados não são eficientes para a classificação do novo caso, pode-se gerar uma nova classificação baseada nos conceitos (classes e instâncias) encontrados no conteúdo do objeto de estudo, bem como por meio dos adjetivos presentes, resultando uma nova proposta de classificação.

A terceira etapa do modelo é opcional, e o fato dela existir ou não caracteriza a implementação como automática ou semiautomática. Nesse estágio, é possível que um especialista interaja com a classificação sugerida pela aplicação.

Por fim, a quarta e última etapa refere-se à retenção, que consiste na formatação das informações de classificação geradas a partir da análise da sentença em casos que possam ser armazenados, recuperados e reutilizados, permitindo assim que sejam utilizadas classificações passadas em novas categorizações.

O próximo capítulo relata os testes de viabilidade do modelo proposto, apresentando um protótipo como implementação proposta. Também são expostas informações sobre a avaliação do modelo desenvolvido bem como a análise dos resultados inicias.

4 EXPERIMENTOS E EVOLUÇÃO DO MODELO

Este capítulo tem como objetivo apresentar toda a evolução do modelo inicial, proposto no Capítulo 3, para chegar ao modelo final da tese. Para tal, uma série de testes (execuções) foi realizada para justificar as alterações no modelo inicialmente proposto.

Primeiramente, é apresentado um recorte do modelo inicial, a fim de construir uma avaliação preliminar da estratégia isolada e, posteriormente, combinada com os elementos que constituem o modelo (RBC, ontologia e construção da árvore de inferência). Para esses experimentos, foi utilizada uma base de revisões (*reviews*) sobre câmeras digitais.

Após a avaliação preliminar, serão apresentados os pontos de evoluções necessários para alcançar os objetivos deste trabalho. Na sequência, um novo experimento será proposto para avaliar e definir métodos para criar e manter atualizada a base de conhecimento. Nesse caso, será utilizada uma base de *reviews* de filmes.

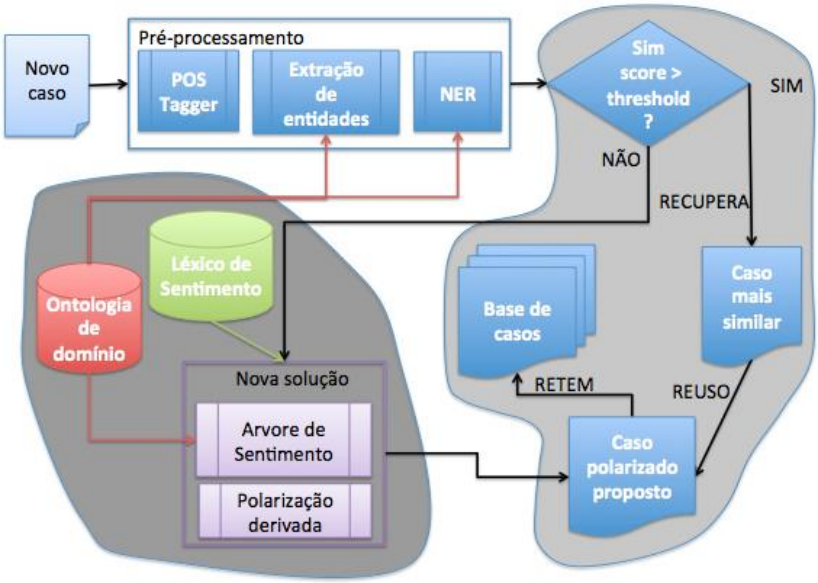
Além disso, serão apresentados métodos para a adequação automática do léxico, o cálculo o *threshold* de um novo domínio, assim como, a melhor estratégia para a tarefa de adaptação. Por fim, é efetuada uma avaliação detalhada do modelo final gerado, seguida pela apresentação e análise dos resultados.

4.1 AVALIAÇÃO PRELIMINAR DO MODELO PROPOSTO

Nesta etapa, optou-se por fazer um recorte do modelo da tese voltado à avaliação do desenvolvimento da árvore de sentimento, considerando os conceitos da ontologia que estão presentes nas sentenças, bem como os adjetivos relacionados e sua polarização, combinando com o uso do RBC.

Para atestar a viabilidade do modelo inicialmente proposto com base no recorte, foi realizada a construção de um estudo visando avaliar a eficiência do seu ponto central, que é a combinação de uma abordagem baseada em caso utilizando etapas do ciclo clássico de RBC, com o uso de uma ontologia de domínio. A Figura 30 demonstra mais detalhes sobre o modelo.

Figura 29 – Modelo adaptado para a avalição de viabilidade



Fonte: Elaborado pelo autor

Inicialmente, o novo caso é submetido ao pré-processamento, no qual seu conteúdo é divido em sentenças. Para cada sentença, todas as palavras são classificadas a partir das suas classes gramáticas utilizando o POS *Tagger*. O próximo passo é a extração de todas as entidades, termos candidatos, presentes na sentença, os quais são submetidos ao processo de reconhecimento de entidades, utilizando-se como base os conceitos da ontologia de domínio e as suas instâncias. A Figura 30 representa um exemplo de saída proveniente da etapa de pré-processamento.

Figura 30 – Exemplo de saída a partir da etapa de pré-processamento

```
SENTENCE => camera has good features and decent battery life for a "regular" camera.
WORDS:
      - camera (Noun, singular or mass) {CONCEPT: camera}
      - has (Verb, 3rd person singular present) [0]
      - good (Adjective) [1]
      - features (Noun, plural) [0]
      - and (Coordinating conjunction) [0]
      - decent (Adjective) [1]
      - battery (Noun, singular or mass) {CONCEPT: battery}
      - life (Noun, singular or mass) [0]
      - for (Preposition or subordinating conjuntion) [0]
      - a (Determiner) [0]
      - "regular" (Noun, singular or mass) [0]
      - camera. (Noun, singular or mass) {CONCEPT: camera}
```

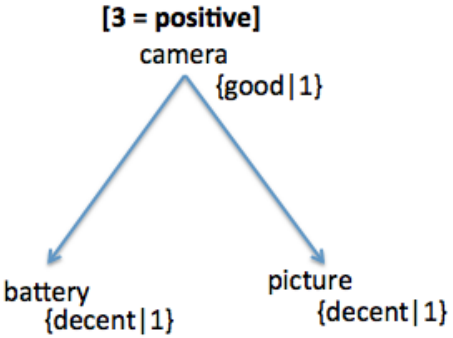
Fonte: Elaborado pelo autor

Pode-se perceber que todas as palavras da sentença foram marcadas com a sua classe gramatical. Para as entidades reconhecidas, foi também anotado o seu conceito relacionado.

Em seguida, são selecionados todos os adjetivos e entidades reconhecidas para serem utilizadas como base na etapa de recuperação. Ao submetê-los, é recuperado o caso mais similar. Se o grau de similaridade for superior ao *threshold*, o caso recuperado é apresentado como possível solução. Quando o caso mais semelhante tiver o seu grau de similaridade (*score*) inferior ao *threshold*, o novo caso é submetido ao processo de nova solução. A partir dos conceitos e das instâncias reconhecidas, é construída uma árvore baseada no caminhamento entre os conceitos da ontologia.

Além disso, são relacionados os adjetivos com a sua polarização (positivo ou negativo). A partir dos adjetivos polarizados, é atribuído o valor 1 para adjetivos marcados como positivos e -1 para adjetivos marcados como negativos. Para ter essa informação da polarização do adjetivo, utiliza-se um léxico de sentimento (ou léxico de opinião, como alguns trabalhos referenciam). O valor final, gerado pelo somatório do grau de cada adjetivo, demonstra se a sentença em questão é positiva ou negativa. A Figura 31 exibe mais detalhes sobre a árvore gerada.

Figura 31 – Exemplo da árvore de sentimento gerada a partir de uma sentença



Fonte: Elaborado pelo autor

No exemplo, apresentado pela Figura 30, pode-se verificar que foi efetuado o somatório do grau dos adjetivos polarizados. Cada adjetivo tem uma ligação direta com uma entidade reconhecida que chega ao valor 3, indicando assim que a sentença é positiva.

A próxima seção traz mais detalhes sobre o cenário de avaliação proposto.

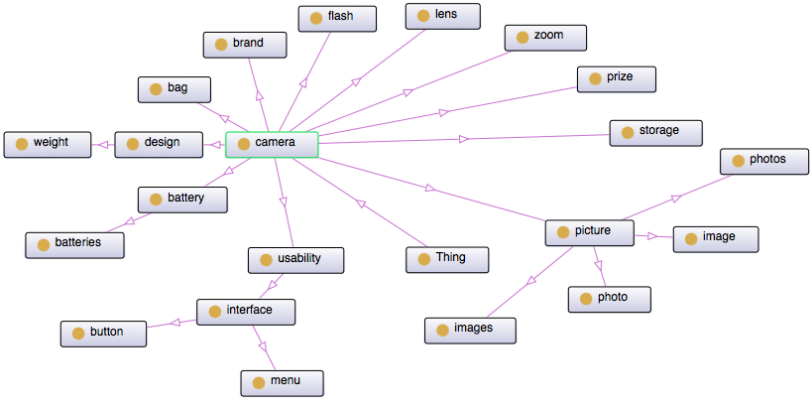
4.1.1 Cenário de avaliação

Foram selecionados 1999 *reviews* relacionados com ao domínio de aplicação de câmeras digitais, os quais foram anotados a partir do *rating*, quantidades de estrelas dadas pelos usuários, como positivos (quando foram marcadas com quatro ou cinco estrelas) ou negativos (quando marcado com uma ou duas estrelas). Esses *reviews* foram extraídos do site da Amazon®[12]. Essa base de validação pode ser obtida a partir do seguinte *dataset*: "unprocessed.tar.gz\sorted_data\camera_&_photo", apresentado com o nome *Multi-DomainSentimentDataset (version 2.0)*, disponível em: http://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

Nessa base de *reviews* da Amazon®, existem dois arquivos: *positive.review* e *negative.review*. No primeiro são apresentados 1000 *reviews* marcados como positivos e, no segundo arquivo, 999 *reviews* marcados como negativos. Esses *reviews* foram extraídos e processados a partir de um documento textual e armazenados em uma base relacional para facilitar os testes posteriores.

Para a construção da ontologia de domínio no contexto de câmeras digitais, utilizou-se como base a ontologia apresentada no trabalho de Wei e Gulla (2010). Vale ressaltar que a ontologia foi estendida para representar melhor o domínio em questão. A Figura 32 apresenta mais detalhes sobre a ontologia estendida.

Figura 32 – Ontologia estendida baseada no trabalho de Wei e Gulla (2010)



Fonte: Elaborado pelo autor

[12] Acesse em www.amazon.com

Tendo como base as classes e as instâncias da ontologia de domínio, optou-se por descartar todos os *reviews* da base que não possuiam relação com os conceitos do domínio, chegando então a uma população de 1443 *reviews*, dos quais 726 (50,3%) positivos e 717 (49,7%) negativos. Na seção a seguir, são apresentados os detalhes de como foi implementado o protótipo de solução para avaliar o modelo proposto.

4.1.2 Protótipo da solução

O protótipo desenvolvido respeita as etapas e os passos propostos no modelo apresentado na Figura 29. Para o pré-processamento das tarefas de quebra da sentença e POS *Tagger*, utilizou-se a biblioteca livre *Apache OpenNLP*[13], que foi construída em Java e possui uma série de algoritmos para auxiliar a área de processamento de linguagem natural.

Para a estratégia de reconhecimento das instâncias e de classes da ontologia no texto, utilizou-se a implementação e o modelo proposto por Ceci et al. (2012).

Para calcular a similaridade dos casos, utilizou-se a biblioteca de indexação de busca *Apache Lucene*[14], também construída em Java e livre. O *Apache Lucene* utiliza como base para o seu cálculo de similaridade o valor do cosseno entre os vetores de documentos.

O processo de polarização dos adjetivos e advérbios foi efetuado a partir da utilização de um léxico de sentimento (ou opinião, como alguns trabalhos o referenciam). Para esta proposta de solução, optou-se pela utilização do léxico apresentado por Hu e Liu (2004). O motivo para a utilização desse léxico é sua simplicidade, permitindo uma implementação fácil e rápida, e devido à grande quantidade de trabalhos que o utilizam, como, por exemplo, Zhang e Liu (2011a), Dong et al. (2013a), Dong et al. (2013b), entre outros.

As demais etapas foram construídas para atender necessidades do modelo, utilizando a linguagem de programação Java.

4.1.3 Experimentos

Tendo em vista a implementação, utilizou-se como população para os testes os 1443 *reviews* sobre câmeras digitais retirados da Amazon®, apresentados na Seção 4.1.1. Todos os *reviews* foram processados e armazenados na base de casos da proposta de solução.

---

[13] Saiba mais sobre a biblioteca livre *Apache OpenNLP* em: http://opennlp.sourceforge.net/
[14] Conheça a biblioteca Lucene em: http://lucene.apache.org

Para avaliação da acurácia do modelo, utilizou-se a técnica chamada *leave-one-out cross validation* (LOOCV), que consiste na retirada de um caso armazenado da base e comparação dele mesmo com os demais casos a partir da recuperação do caso mais similar. São comparadas as classificações do caso retirado com as do caso recuperado. Se as duas forem iguais, considera-se que a classificação obteve suceso, caso contrário, não (DEVROYE e WAGNER, 1979). Essa primeira natureza de teste procura avaliar a estratégia de recuperação dos casos baseando-se na utilização do RBC.

A partir de uma sentença, pretende-se reconhecer as suas entidades e, com base nelas, busca-se o caso mais similar, ou seja, o caso que apresentar o maior grau de similaridade tem a sua polarização apresentada como possível solução. Para esse teste, formularam-se quatro execuções com características distintas, quais sejam:

- Execução 1: a partir de uma sentença reconhece-se todos os conceitos e instâncias da ontologia que estão presentes. Após o reconhecimento, busca-se por todos os casos que possuem **todos** os conceitos em seu texto.
- Execução 2: utilizando uma sentença, são reconhecidos os conceitos e as instâncias da ontologia. Na sequência, são recuperados casos que possuem **pelo menos** um conceito ou uma instância da ontologia de domínio.
- Execução 3: são utilizados todos os critérios da Execução 1, mas também são considerados os **adjetivos** reconhecidos na sentença para a recuperação dos casos.
- Execução 4: são considerados todos os critérios da Execução 2, mas são igualmente levados em consideração os **adjetivos** reconhecidos na sentença para a recuperação dos casos.

Os resultados obtidos nas execuções estão apresentados na Tabela 4.

Tabela 4 – Execução inicial do protótipo de solução

Execução	Acertos	Erros	Taxa de acertos
1	738	705	51,1%
2	743	700	51,4%
3	149	1294	10,3%
4	1030	413	**71,4%**

Fonte: Elaborado pelo autor

A partir da Execução 4, que obteve a melhor taxa de acerto, levou-se em consideração o maior grau de similaridade dos termos

recuperados para cada busca na base de casos. O valor médio para o *threshold* encontrado foi de 0,43.

Pode-se perceber que a execução que obteve menor acerto foi a Execução 3. O motivo para tal resultado é que, nessa execução, são utilizadas todas as instâncias, conceitos e termos polares encontrados, de modo que só será considerado um caso similar se ele apresentar todos os termos em questão, o que dificulta, e muito, acertar o caso.

A segunda natureza de testes teve o intuito de identificar a estratégia de geração da árvore de sentimento. Para este caso, utilizou-se como base os *reviews* com conceitos ou instâncias vindas da ontologia em seu texto (1443 registros). Para cada *review*, foi gerada uma árvore de inferência baseada nos seus conceitos e nos adjetivos relacionados com o objetivo de gerar de uma polarização. O próximo passo consiste na comparação da polaridade inferida com a classificada originalmente. O resultado obtido está apresentado na Tabela 5.

Tabela 5 – Testes utilizando a geração da árvore de sentimento

População	**Acertos**	**Erros**	**Taxa de acerto**
1443	989	454	68,5%

Fonte: Elaborado pelo autor

O terceiro tipo de teste busca executar as duas soluções de maneira híbrida, ou seja, utilizando RBC e, quando necessário, gerando a árvore de inferência.

Todos os 1443 registros previamente polarizados (base anotada) foram submetidos ao processo de geração da árvore de inferência. Ao final do processo, todos os *reviews* que tiveram sua polarização classificada corretamente em relação à polarização previamente anotada foram armazenados como eventos da base de casos. A partir de uma sentença, pretende-se reconhecer as suas entidades e buscar o caso mais similar e, com base no caso recuperado, será utilizada a sua polarização como possível solução. Se o grau de similaridade for maior que o *threshold* definido, o caso recuperado é apresentado como solução, caso seja inferior, é gerada a árvore de inferência com uma nova proposta de polarização.

A média do grau de similaridade dos testes executados, utilizando como base a Execução 1, foi de 0,4. A Tabela 6 apresenta o resultado obtido.

Tabela 6 - Teste utilizando a abordagem híbrida (RBC + árvore de sentimento)

População	**Acertos**	**Erros**	**Taxa de acerto**
1443	1100	343	76,2%

Fonte: Elaborado pelo autor

Para obter uma melhor classificação, buscou-se identificar os erros de categorização, utilizando como base os 343 casos qualificados de maneira incorreta. A próxima seção apresenta mais detalhes sobre essa etapa.

4.1.3.1 Melhorias no protótipo

A partir da análise dos resultados anteriores, identificou-se os fatores de sucesso bem como os de insucesso. A qualidade dos resultados origina-se não somente por meio da reutilização de casos, como também pela construção das árvores.

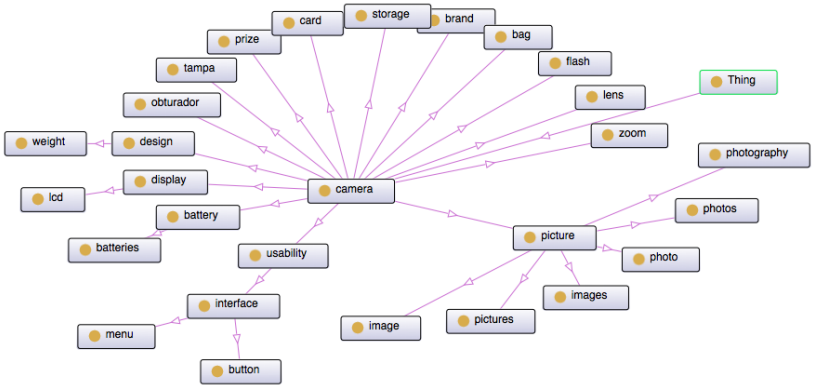
A análise dos fatores que levaram a polarizações erradas é a base das melhorias apresentadas nesta seção. Como já foi amplamente discutido neste trabalho e, particularmente, na literatura sobre análise de sentimento, fatores de fracasso originam-se do não tratamento da negação, da ambiguidade e da subjetividade.

Primeiramente, optou-se por tratar os problemas relacionados à ambiguidade. Para fazer a análise dos casos com erro, 80 casos foram selecionados, de maneira aleatória, dentre os 383 com problemas de polarização. Desses 80 casos, metade deveria ser positiva e a outra metade, negativa.

Verificou-se que muitos dos termos utilizados no léxico de sentimento proposto no trabalho de Hu e Liu (2004) possuíam polarização errada tendo em vista o domínio de aplicação. Por exemplo, segundo o léxico, o termo *intense* tinha uma conotação negativa, mas no domínio de aplicação, isso não é verdade. Foram retirados basicamente 23 termos das listas de positivos e negativos e adicionados dois novos termos, um para a lista de palavras positivas e outro para a lista de palavras negativas. Após a manutenção do léxico de sentimento, fez-se o mesmo com a ontologia de domínio, adicionando nove novas classes (conceitos) à ontologia[15] de domínio. A Figura 33 apresenta a ontologia expandida.

[15] A ontologia de câmera possue as ligações no formato de “é um” para demonstrar hierarquia entre classes. Adotou-se esse formato para respeitar o modelo proposto por Wang, Nie e Liu (2013).

Figura 33 – Ontologia expandida



Fonte: Elaborado pelo autor

As modificações detalhadas acima podem ser analisadas na Tabela 7. É importante mencionar que o resultado obtido pelo tratamento Manutenção do léxico de sentimento foi a partir da estratégia descrita na Tabela 6, já a execução do tratamento da Manutenção na ontologia de domínio deu-se a partir da Manutenção do léxico de sentimento, ou seja, a taxa de acerto está aumentando à medida que se adiciona novos tratamentos.

Tabela 7 - Melhoria da taxa de acerto a partir da manutenção da base de conhecimento

Tratamento	**População**	**Acerto**	**Taxa de acerto**
Manutenção do léxico de sentimento	1443	1120	77,6%
Manutenção na ontologia de domínio	1443	1193	82,6%

Fonte: Elaborado pelo autor

Conforme foi apresentado na revisão da literatura, quando existe uma negação na sentença, a polarização do termo em questão deve ser alterada. Para tratar essa situação, optou-se por desenvolver um algoritmo que, ao encontrar algum termo que tenha como característica a negação, analisa uma janela – 4 termos antes e 4 termos depois – para verificar se existe um termo polarizado. Caso exista, é alterada a polarização do termo mais próximo à negação. O resultado dessa abordagem pode ser verificado na Tabela 8. É importante mencionar que

esse resultado foi obtido a partir dos resultados do tratamento Manutenção na ontologia de domínio, ou seja, de maneira incremental.

Tabela 8 - Teste com tratamento de negação

Tratamento	População	Acerto	Taxa de acerto
Negação	1443	1260	87,3%

Fonte: Elaborado pelo autor

Por fim, foi realizado o tratamento de problemas envolvendo a subjetividade. Verificou-se que muitos dos *reviews* não estavam relacionados propriamente com o domínio em questão, nesse caso, câmera digital. Muitos falavam sobre como estava o dia ou como foi uma viagem. Tendo em vista que o tratamento de subjetividade é uma área complexa e isso foge do escopo deste trabalho, optou-se por utilizar como base para validação apenas os *reviews*, da população de 1443, que possuíam o *rating* (classificação atribuída pelo revisor) 1 (negativo) ou 5 (positivo). Nesse caso, a população passou a ser de 942 registros. O resultado é apresentado na Tabela 9.

Tabela 9 – Teste com os *reviews* que possuem *rating* 1 ou 5

Tratamento	População	Acerto	Taxa de acerto
Utilização rating 1 ou 5	942	852	90,4%

Fonte: Elaborado pelo autor

Percebe-se que, ao trabalhar com uma base em que a polarização positiva e negativa está bem definida, ou seja, não existe subjetividade, a execução do protótipo alcança uma taxa de acerto ainda maior em relação aos demais testes, ultrapassando os 90%. Isso leva a crer que a aplicação da solução em um ambiente em que as sentenças sejam melhor avaliadas, anotadas manualmente para serem utilizadas como teste, pode-se chegar a uma taxa melhor.

A seção a seguir tem como objetivo definir algumas métricas de avaliação e demostrar a significância estatística dos testes executados.

4.1.3.2 Significância estatística e métricas de avaliação

Para definir quais métricas de avaliação seriam aplicadas neste trabalho, foram analisados artigos relacionados ao tema e à abordagem proposta. A partir da análise dos trabalhos de Hu e Liu (2004), Kazama e Tsujii (2005), Zhang e Liu (2011), Moreo et al. (2012), Cruz et al. (2013), Moraes et al. (2013) e Li e Tsai (2013), identificou-se que as

medidas mais utilizadas para a avaliação desse contexto são: acurácia, precisão (*precision*), revocação (*recall*) e medida F (*f-measure*).

Para calcular tais medidas, é necessária a utilização de uma matriz de contingência, apresentada na Figura 34 com mais detalhes.

Figura 34 – Matriz de contingência

		Classificação correta (esperada)	
		Positivo	Negativo
Classificação proposta	Positivo	TP (positivo verdadeiro)	FP (falso positivo)
	Negativo	FN (falso negativo)	TN (verdadeiro negativo)

Fonte: Elaborado pelo autor

Onde,

- **TP** (positivo verdadeiro): casos positivos classificados corretamente;
- **TN** (verdadeiro negativo): casos negativos classificados corretamente;
- **FP** (falso positivo): casos positivos classificados erroneamente, ou seja, deveriam ser negativos, mas foram classificados como positivos; e
- **FN** (falso negativo): casos negativos classificados erroneamente, ou seja, deveriam ser positivos, mas foram classificados como negativos.

Para calcular a acurácia, utiliza-se a Equação 12.

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

A precisão é obtida por meio da aplicação da Equação 13.

$$Precisão = \frac{TP}{TP + FP} \tag{13}$$

A revocação pode ser calculada a partir da Equação 14.

$$\text{Revocação} = \frac{TP}{TP + FN} \tag{14}$$

Para calcular a medida F, pode-se adotar a Equação 15.

$$\text{Medida F} = 2 . \frac{\text{precisão} . \text{revocação}}{\text{precisão} + \text{revocação}} \quad (15)$$

O método escolhido para verificar a significância estatística de uma alteração ou evolução do modelo é o teste de McNemar, o qual pode ser utilizado para atestar a significância estatística de mudanças, ou seja, é aplicável para situações do tipo antes e depois de algo (MCNEMAR, 1947).

No contexto desta tese, ao afirmar que duas execuções possuem significância estatística, quer-se dizer que elas não ocorrem de maneira igual e que seus resultados não foram obtidos pelo acaso. Para o cálculo de McNemar, utilizando como base dois testes quaisquer, t_1 e t_2, é necessário coletar as variáveis c_{01} e c_{10}. A seguir são definidas as variáveis.

- O número de instâncias classificadas erroneamente por t_1 e t_2 (c_{00});
- O número de instâncias classificadas erroneamente por t_1, mas classificadas corretamente por t_2 (c_{01});
- O número de instâncias classificadas erroneamente por t_2, mas classificadas corretamente por t_1 (c_{10});
- O número de instâncias classificadas corretamente por t_1 e t_2 (c_{11}).

Com base nessa definição, o teste de McNemar pode ser obtido pela Equação 16, em que $c_{01} + c_{10} >= 20$.

$$x^2 = \frac{(|c_{01} - c_{10}| - 1)^2}{c_{01} + c_{10}} \quad (16)$$

O teste de McNemar é interpretado como uma tabela Chi-Quadrado padrão. Como é utilizada, neste estudo, uma classificação binária (positivo ou negativo), apenas a primeira linha da tabela de contingência é usada, designada para um grau de liberdade. Para um grau de confiança de 95% ($p < 0,05$), o resultado do cálculo só será considerado se o valor de McNemar for maior que 3,841.

O teste de McNemar é utilizado na tese de duas maneiras diferentes:

(1) Para verificar se, ao adicionar uma nova característica (etapa) ao modelo, ela realmente tem uma contribuição válida para as classificações. Nesse caso, por meio do valor de McNemar, percebe-se se existe ou não significância estatística; e

(2) Para comparar a execução do modelo proposto com outro método e/ou técnica. Nesse caso, é verificado se existe significância estatística entre as execuções, e se os valores das colunas c_{01} e c_{10} podem auxiliar na análise dos testes.

A partir das métricas apresentadas nesta seção juntamente com o teste de McNemar, define-se então o método de avaliação do modelo proposto a fim de atestar sua viabilidade.

4.1.3.3 Análise estatística sobre os testes iniciais

Utilizando como base os testes iniciais efetuados a partir da proposta de solução, são aplicados dois grupos de testes.

O primeiro grupo de teste compara a evolução do modelo considerando a população de 1443 casos. No primeiro teste, leva-se em conta a aplicação-base (apresentada na Tabela 6) e, no segundo teste, considera-se a situação após as melhorias propostas. A Tabela 10, a seguir, apresenta essa comparação.

Tabela 10 – Medidas de avaliação do protótipo-base e melhorado

Teste$_{1443}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base	655	468	76	244	0,78	0,90	0,73	0,80
Melhorado	673	588	53	129	0,87	0,93	0,84	0,88
%Variação	2,75	25,6	-30	-47	11,5	3,3	15	10

Fonte: Elaborado pelo autor

Os rótulos das colunas, “Acu.”, “Prec.”, “Rec.” e “F”, referem-se a acurácia, precisão, revocação e medida F, respectivamente. A última linha da tabela demonstra a variação percentual, considerando o teste-base e o teste melhorado, e apresenta um aumento de 9% na acurácia.

O segundo grupo de testes utilizou como base os 942 *reviews*, que possuem um *rating* 1 ou 5, extraídos dos 1443. A Tabela 11, a seguir, apresenta a comparação entre os valores.

Tabela 11 – Medidas de avaliação sobre a base de 942 *reviews*

Teste$_{942}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base	460	287	50	145	0,79	0,90	0,76	0,83
Melhorado	481	373	25	63	0,91	0,95	0,88	0,92
%Variação	4,6	30	-50	-56	15	5,6	15,8	10,8

Fonte: Elaborado pelo autor

Na seção seguinte, apresenta-se a aplicação de dois dos principais algoritmos utilizados que obtiveram melhores resultados no contexto da análise de sentimento: SVM e *Naïve Bayes*.

4.1.3.4 Análise comparativa

Para a comparação detalhada desta seção, utilizou-se os algoritmos SVM e *NaiveBayes* (NB) com os mesmos dados aplicados nos testes apresentados na Seção 4.1.3.3.

A modelagem e a análise foram realizadas utilizando *LightSide*[16], uma ferramenta de mineração de texto e aprendizagem de máquina. *LightSide* é uma ferramenta de código aberto, e utiliza como base as bibliotecas de código aberto, *Weka*[17]e *LibLinear*[18], para sua fundação computacional. A Tabela 12 demonstra os resultados obtidos.

Tabela 12 – Testes utilizando SVM e NB sobre a base de *reviews*

Teste	TP	TN	FP	FN	Acu.	Prec.	Rec.	F
NB_{1443}	620	600	106	117	0,84	0,85	0,84	0,84
SVM_{1443}	595	625	131	92	0,84	0,81	0,86	0,84
NB_{942}	430	389	76	47	0,86	0,84	0,90	0,87
SVM_{942}	442	382	64	54	0,87	0,87	0,89	0,88

Fonte: Elaborado pelo autor

Como apresentado na Tabela 12, foram calculados os valores para cada instância do conjunto de dados com intuito de determinar a significância estatística em relação à abordagem proposta. A maior precisão foi de 87% usando SVM para o conjunto de dados com 942 casos. Essa mesma precisão foi alcançada com a abordagem melhorada (ver Tabela 8 para o conjunto de 1443 comentários de dados). Ainda para o conjunto de dados com 942 casos, a abordagem melhorada atingiu uma acurária de 90,4% (ver Tabela 9). A Tabela 13 fornece o valor de McNemar mostrando sua significância estatística.

Tabela 13 – Significância estatística entre o modelo proposto e os demais modelos

Método 1	Método 2	c_{01}	c_{10}	McNemar
$Tese_{1443}$	NB_{1443}	127	169	5,68
$Tese_{1443}$	SVM_{1443}	120	162	5,96
$Tese_{942}$	NB_{942}	65	101	7,38
$Tese_{942}$	SVM_{942}	60	91	5,96

Fonte: Elaborado pelo autor

A partir de todos os dados aqui expostos, percebe-se que a abordagem proposta possui significância estatística quando comparada

[16]Saiba mais sobre LightSide em: http://lightsidelabs.com
[17]Conheça a Weka em: http://www.cs.waikato.ac.nz/ml/weka/
[18]Mais detalhes sobre a biblioteva LibLinear em: http://www.csie.ntu.edu.tw/~cjlin/liblinear/

às duas outras abordagens (SVM e NB). Na próxima seção, objetiva-se identificar os próximos passos para esta pesquisa.

4.2 EVOLUÇÃO DO MODELO PROPOSTO

Esta seção tem como objetivo apresentar as evoluções no modelo proposto a partir das conclusões obtidas no experimento inicial. Percebeu-se que as alterações efetuadas no léxico de sentimento apresentaram uma solução para o domínio de aplicação específico, mas essas alterações podem apresentar problemas quando aplicadas em outro domínio. A próxima seção descreve as alterações necessárias na ontologia de domínio para suportar as particularidades de cada contexto de aplicação.

4.2.1 Evolução da ontologia de domínio

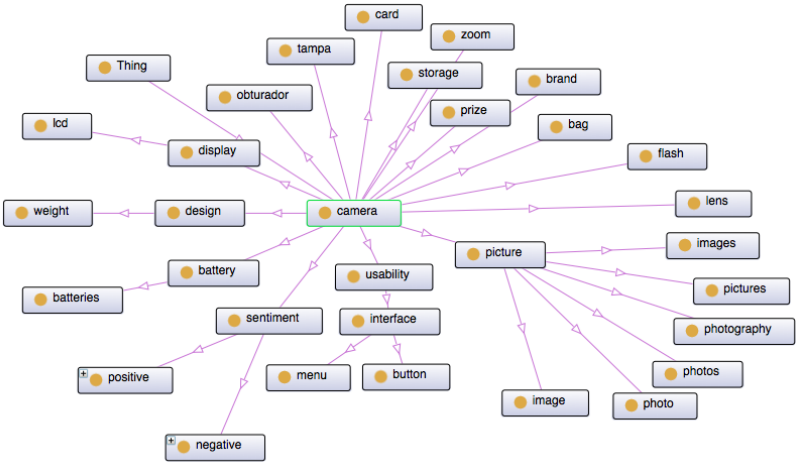
Durante o experimento inicial efetuado no modelo proposto, foi necessário alterar alguns termos do léxico de sentimento sugerido por Hu e Liu (2004). Percebeu-se que esses termos têm conotações distintas dependendo do domínio. Como existe apenas um léxico de sentimento, ao alterar a polarização de um termo, todos os domínios que fizerem uso desse léxico teriam impacto na sua classificação.

Optou-se por armazenar as alterações necessárias no léxico de sentimento diretamente na ontologia de domínio. Para isso, foi necessário criar mais três classes (conceitos). Essas três classes devem ser incorporadas em todas as ontologias de domínio que forem utilizadas pelo modelo proposto. A seguir são apresentadas as novas classes.

- *Sentiment*: classe que deve ser adicionada ao conceito principal da ontologia de domínio, ou seja, filha do conceito-chave;
- *Positive*: classe-filha de *Sentiment*. Deve abrigar todas as instâncias que forem alteradas no léxico de sentimento relacionadas aos termos positivos; e
- *Negative*: também filha de *Sentiment*, abriga as instâncias que foram alteradas no léxico de sentimento relacionadas aos termos negativos.

A Figura 35, a seguir, apresenta a ontologia de domínio de câmera, utilizada na Seção 4.1, como base para os experimentos iniciais, contemplando as novas classes.

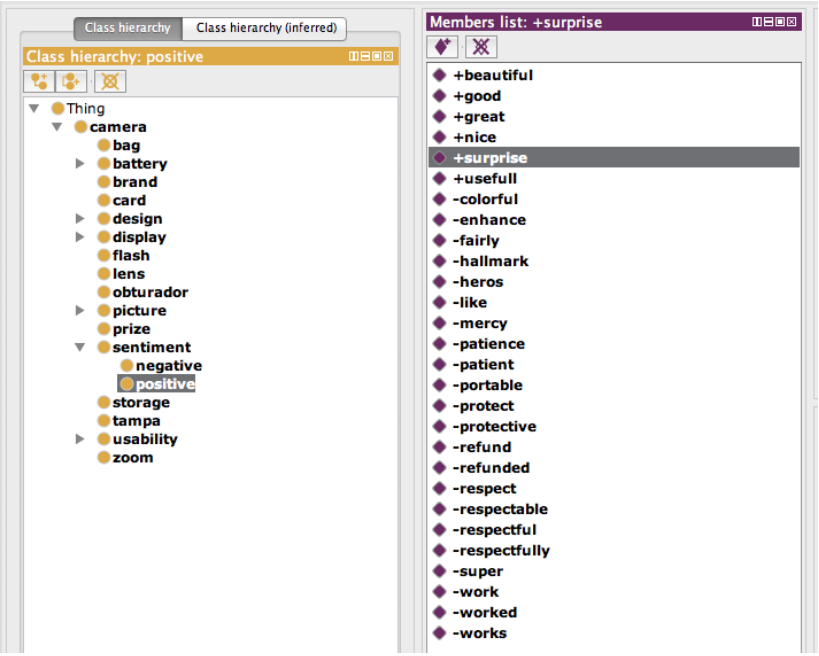
Figura 35 – Ontologia de câmera com as novas classes



Fonte: Elaborado pelo autor

Como afirmado anteriormente, todos os termos que foram retirados ou adicionados de uma classe do léxico de sentimento (positivo ou negativo) foram adicionados na ontologia de domínio. A Figura 36 apresenta como os termos retirados ou adicionados ao léxico ficaram organizados na ontologia de domínio na classe *Positive*.

Figura 36 – Termos do léxico vinculados à classe *Positive* da ontologia

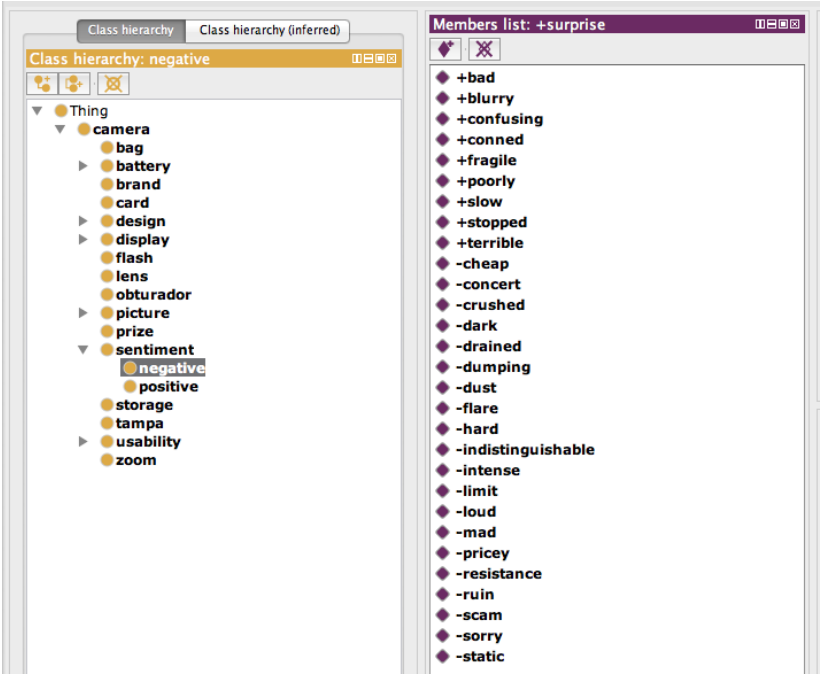


Fonte: Elaborado pelo autor

Pode-se perceber que, para cada termo anteriormente retirado do léxico de sentimento, foi adicionado um termo como instância da classe *Positive*, com o sinal "-" (menos) na frente do nome da instância. No caso dos termos que foram adicionados ao léxico, aparecem também como instâncias da classe *Positive*, mas neste caso, com o sinal "+" (mais) no início.

Esse mesmo processo foi efetuado para os termos que sofreram alterações do léxico de sentimento relacionados à lista de termos negativos. A Figura 37 apresenta essa relação.

Figura 37 – Termos do léxico vinculados à classe *Negative* da ontologia



Fonte: Elaborado pelo autor

As modificações efetuadas na ontologia de domínio permitiram atingir a mesma taxa de acerto dos experimentos anteriores, o que demonstra que é possível utilizar essa prática para manter as especificidades do domínio diretamente na ontologia.

A próxima seção tem como objetivo relatar a aplicação do modelo já contemplando as modificações executadas na ontologia de domínio, em um novo contexto de aplicação. Objetiva-se também, para a próxima seção, a formalização de um método para aplicar o modelo em um novo domínio.

4.2.2 Aplicação do modelo no domínio de filmes

O modelo proposto neste trabalho foi inicialmente aplicado sobre uma base de *reviews* construídos pelos clientes da organização de comércio eletrônico Amazon®. O domínio escolhido foi o de câmeras digitais. Nesta seção, objetiva-se utilizar a mesma natureza de *reviews* sobre produtos da Amazon®, mas, nesse caso, o domínio de aplicação será o de filmes vendidos na forma de DVDs.

A base de *reviews* utilizada foi obtida por meio do mesmo pacote descrito na Seção 4.1.1, ou seja, a base de validação foi obtida pelo seguinte *dataset*: unprocessed.tar.gz\sorted_data\dvd, apresentado com o nome *Multi-DomainSentimentDataset (version 2.0)*, disponível em http://www.cs.jhu.edu/~mdredze/datasets/sentiment/.

Os arquivos utilizados foram o *negative.reviews* que contém os *reviews* classificados como negativos e o *positive.reviews*, com os *reviews* classificados como positivos. O arquivo que contém os *reviews* positivos possui 996 instâncias, e o arquivo que possui os *reviews* negativos possui 995 instâncias.

O próximo passo, após definir a base de *reviews*, envolve a construção ou a seleção de uma ontologia de domínio. A subseção a seguir apresenta mais detalhes sobre esse recurso.

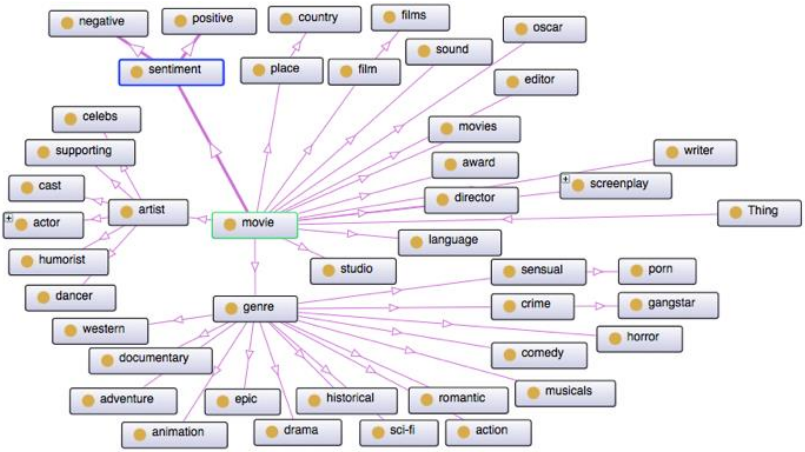
4.2.2.1 Ontologia do domínio de filmes

Nesta etapa, foi utilizada uma ontologia do domínio de filmes, *Movie Ontology*, obtida a partir do *link* http://www.movieontology.org/. A ontologia foi construída por Amancio Bouza do Departamento de Informática da Universidade de Zurich. Disponibilizada em 2010, já foi utilizada em alguns trabalhos da área de análise de sentimento, como, por exemplo, Rahayu et al. (2010a), Rahayu et al. (2010b), Freitas e Vieira (2013a) e Peñalver-Martinez (2014).

A *Movie Ontology* possui muitos conceitos herdados de recursos da *Web*, sem aplicabilidade no contexto do trabalho e no domínio de aplicação. Nesse sentido, optou-se por eliminar conceitos que não fazem parte do domínio, como, por exemplo, a classe Profissões, que apresenta conceitos como bombeiro, eletricista, entre outros. Assim como na ontologia de câmera, essa ontologia foi estendida para suportar a classe *Sentiment* e as suas subclasses *Positive* e *Negative*. A Figura 38 apresenta a ontologia de domínio final[19] gerada a partir da *Movie Ontology*.

Figura 38 – Ontologia gerada a partir da *Movie Ontology*

[19] A ontologia de filme possue as ligações no formato “é um” para demonstrar hierarquia entre classes. Adotou-se esse formato para respeitar o modelo proposto por Wang, Nie e Liu (2013).



Fonte: Elaborado pelo autor

A partir da base de *reviews* sobre filmes, verificou-se quantas revisões possuem conceitos da ontologia. Todos os *reviews* possuem pelo menos um conceito da ontologia.

Uma vez já selecionada a base de avaliação, que pode ser considerada como base de treinamento, e com a ontologia de domínio construída, faz-se necessário definir o valor do *threshold* que vai parametrizar se um caso recuperado pode ser considerado como proposta de solução ou não para o novo problema (nova sentença a ser polarizada). A seção a seguir apresenta mais detalhes sobre a definição do *threshold*.

4.2.2.2 Método para calcular o *threshold*

O *threshold* é utilizado para definir se um caso recuperado (proposta de solução) é válido ou não para a nova situação problema (nova sentença a ser polarizada).

Para chegar a um valor adequado para o processo de recuperação de casos conhecidos (árvores de inferências já geradas para sentenças passadas), inicialmente, deve-se ter uma base de casos já armazenados, os quais podem ter sido armazenados durante o processo de polarização ou serem frutos de uma base de treinamento. Deve-se submeter, na sequência, a base inicial coletada ou gerada ao processo de *leave one out cross validation*, que consiste em retirar um caso armazenado e submetê-lo ao processo de busca de um outro caso similar.

Para cada caso, é recuperado o caso mais similar, recuperando o grau de similaridade entre os dois casos. Dessa forma, ao final do

processo de *leave one out*, é possível extrair a média. O valor médio dos graus de similaridade define o valor inicial do *threshold*.

A seção a seguir apresenta os passos necessários para aplicar o modelo proposto a um novo domínio de aplicação, utilizando como base os elementos já mencionados neste capítulo.

4.2.2.3 Passos necessários para aplicar o modelo proposto a um novo domínio

Para aplicar o modelo proposto, deve-se seguir alguns passos, que são apresentados na sequência.

- **Passo 1**: identificar o conceito-chave do novo domínio.
- **Passo 2**: buscar uma ontologia de domínio ou construir uma nova ontologia.
    - Para a construção da ontologia, sem a utilização de outra, pode-se aplicar metodologias ou ferramentas como o OntoKEM (RAUTENBERG et al., 2008).
- **Passo 3**: caso exista uma base para treinamento, esta deve ser submetida ao cálculo do *threshold*.
- **Passo 4**: configurar um léxico de sentimento.
    - Utiliza-se como padrão o proposto por Hu e Liu (2004).
- **Passo 5**: executar o protótipo sobre a base de treinamento ou disponibilizar para o uso.
- **Passo 6**: após a execução do Passo 5, pode-se executar a manutenção e a evolução da base de conhecimento, mais detalhes sobre essa etapa são apresentados na Seção 4.2.3.

A seção a seguir apresenta a aplicação dos passos descritos implementados em um novo domínio de aplicação, filmes.

4.2.2.4 Aplicação do modelo para o domínio filmes

Inicialmente, conforme solicitado no Passo 1, apresentado na seção anterior, optou-se pela definição do conceito-chave do domínio de aplicação. Como a base de *reviews* coletados foi retirada da Amazon.com americana, os *reviews* estão escritos em inglês. Em função disso, o conceito-chave definido foi *movie*.

O Passo 2 foi apresentado com mais detalhes na Seção 4.2.2.1, no qual se utilizou como base um recorte da *Movie Ontology*, conforme apresentado na Figura 38.

O Passo 3 consiste na utilização da base de treinamento, no caso foram utilizados os 1991 *reviews* recuperados da Amazon®, para calcular o valor do *threshold*. Utilizando como base o método definido na Seção 4.2.2.2, o valor do *threshold* foi de 0.18. Este *threshold* será utilizado nos experimentos das próximas seções.

O Passo 4 consiste na utilização de um léxico de sentimento como base. Conforme já utilizado nos demais experimentos, optou-se pela aplicação do léxico proposto por Hu e Liu (2004).

O Passo 5 tem como objetivo executar o protótipo utilizando os elementos descritos nos passos anteriores. O resultado obtido está apresentado na Figura 39.

Figura 39 – Console de saída do protótipo de solução do modelo proposto

```
============================================================
Recuperados com polarização correta: 1429
Recuperados com problema de polarização: 562
Recuperados a partir da base de caso: 477
Recuperados a partir da base de caso e utilizado com sucesso: 334
Diferença entre polarização recuperada e gerada (arvore de sentimento): 626
        Arvore com acerto: 399 CBR com acerto: 227
Total de casos processados: 1991
71% de acerto
============================================================
```

Fonte: Elaborado pelo autor

A próxima seção apresenta com mais detalhes os resultados obtidos na execução do protótipo desenvolvido a partir do modelo proposto neste novo cenário de aplicação.

4.2.2.5 Avaliação dos resultados iniciais no domínio de filmes

Utilizando como configuração inicial os itens explicados na Seção 4.2.2.4, chegou-se a uma acurácia de 0.72. A Tabela 14 apresenta mais detalhes sobre os valores encontrados para as medidas precisão, revocação e medida F.

Tabela 14 – Medidas de avaliação do protótipo base no domínio de filmes

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base	771	658	225	337	0,72	0,77	0,70	0,73

Fonte: Elaborado pelo autor

A partir dos resultados obtidos inicialmente, deve-se submeter os casos que foram classificados de maneira errada como entrada para o processo de manutenção e de evolução da base de conhecimento. A Seção 4.2.3 apresenta mais detalhes sobre os passos necessários para adequar o léxico e a ontologia ao domínio de aplicação.

4.2.3 Manutenção e evolução da base de conhecimento

Como base para todo novo domínio de aplicação, considerando o modelo proposto, foi utilizado o léxico de sentimento proposto por Hu e Liu (2004). Percebeu-se, com o primeiro experimento, demonstrado na Seção 4.1.1, que existem particularidades em aplicações de orientações semânticas por domínios de aplicação.

A solução empregada aqui consiste, portanto, na manutenção do léxico, persistindo as particularidades de cada domínio em sua respectiva ontologia, como é apresentado na Seção 4.2.1.

Questões de subjetividades e ambiguidades podem ser melhor tratadas com a manutenção no léxico de sentimento e, principalmente, na ontologia de domínio. A seguir, são apresentados alguns passos para auxiliar na manutenção e na evolução do léxico de sentimento e da ontologia de domínio, referenciada neste trabalho como base de conhecimento.

- **Passo 1**: deve-se selecionar todos os casos que apresentaram polarização errada.
- **Passo 2**: para cada caso deve-se analisar a árvore de sentimento gerada.
- **Passo 3**: deve-se verificar se todos os conceitos estão sendo contemplados na árvore.
    - Caso não estejam, deve-se adicioná-los como conceito à ontologia, respeitando a hierarquia das classes.
- **Passo 4**: depois de verificar se todos os conceitos estão presentes no novo problema (sentença a ser polarizada), deve-se analisar os termos polarizados e a sua orientação semântica (positiva ou negativa).
- **Passo 5**: caso o termo não possua a polarização adequada, deve-se adicioná-lo nas subclasses referentes ao conceito *Sentiment* da ontologia do domínio em questão.
- **Passo 6**: também deve-se verificar se existe algum termo negativo (negação) e, se esse termo está invertendo a polarização.
- **Passo 7**: quando todos os casos tiverem sido analisados, finaliza-se o processo.
    - Pode-se eleger uma amostra para a manutenção da base de conhecimento, sem a necessidade de avaliação de todos os casos problemáticos.

Utilizando os resultados obtidos na seção anterior, foram aplicados os passos descritos nos casos que não obtiveram sucesso a fim de melhorar o resultado e tornar a base de conhecimento mais sincronizada com os elementos do domínio de aplicação.

A Tabela 15 apresenta mais detalhes sobre o resultado do segundo experimento a partir da manutenção da base de conhecimento.

Tabela 15 – Medidas do protótipo base e com a manutenção da base de conhecimento

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base	771	658	225	337	0,72	0,77	0,70	0,73
Manutenção	804	834	192	161	0,82	0,81	0,83	0,82

Fonte: Elaborado pelo autor

A partir da manutenção da base de conhecimento utilizando como orientação o método apresentado, chegou-se a uma melhoria na taxa de acerto em 10%, como pode ser observado pela evolução da acurácia de 0,72 (método-base) para 0,82 (base do conhecimento com manutenção).

A seção seguinte apresenta a aplicação dos algoritmos NB e SVM sobre a base de casos para que seja possível comparar os resultados dessas abordagens.

4.2.4 Comparação dos resultados do modelo, NB e SVM

Para atestar a viabilidade do modelo proposto em um novo cenário de aplicação, objetiva-se a comparação dos resultados obtidos e descritos na seção anterior com a aplicação dos algoritmos de NB e SVM sobre o conjunto de dados de filmes. O resultado obtido da aplicação de tais algoritmos sobre a base de casos em questão está apresentado com mais detalhes na Tabela 16.

Tabela 16 – Resultados obtidos da aplicação do NB e SVM na base de *reviews* sobre filme

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
NB	822	744	174	251	0,79	0,82	0,77	0,79
SVM	785	779	216	211	0,78	0,79	0,78	0,78
Modelo	804	834	192	161	**0,82**	0,81	0,83	0,82

Fonte: Elaborado pelo autor

Pode-se perceber que o modelo proposto obteve a acurácia mais alta com 0,82, seguida pela acurácia do NB 0,79 e, por último, a acurácia do SVM, igual a 0,78. A Tabela 17 apresenta a comparação entre os modelos considerando McNemar.

Tabela 17 – Significância estatística entre o modelo proposto e demais modelos no contexto de filmes

Método 1	Método 2	c_{01}	c_{10}	McNemar
NB	**SVM**	214	221	0,08
SVM	**Modelo**	312	238	10,23
NB	**Modelo**	303	236	8,58

Fonte: Elaborado pelo autor

Como pode ser observado nos valores apresentados na Tabela 16, o modelo proposto apresenta melhores resultados, baseando-se na acurácia obtida, do que as técnicas NB e SVM. Também é possível notar que as execuções entre o modelo e as técnicas apresentam significância estatística, observando a Tabela 17.

Tendo em vista os resultados alcançados pelo modelo proposto, objetiva-se construir um algoritmo que facilite o refinamento do léxico para a adequação da base de conhecimento ao seu domínio de aplicação. A seção a seguir apresenta mais detalhes sobre a implementação do algoritmo para que seja possível automatizar o processo mencionado.

4.2.5 Refinamento automático do léxico

Para facilitar o processo de refinamento da base de conhecimento e, principalmente, para a adequação do léxico utilizado para um domínio de aplicação, optou-se pelo desenvolvimento de um método automático que utiliza como base casos que apresentam problemas de polarização, ou mesmo, uma base mínima de treinamento.

O método proposto pode ser utilizado tanto para a adequação de uma base de conhecimento já constituída, a fim de adequá-la para a natureza dos casos, ou ainda para um domínio em que se tenha apenas o léxico original, proposto por Hu e Liu (2004). A seguir são apresentados os passos do método desenvolvido.

- **Passo 1**: selecionar os casos de entrada para o processo de adequação.
- **Passo 2**: para cada caso, selecionar os termos polarizados (presentes no léxico, juntamente com sua polaridade).
- **Passo 3**: para cada termo polarizado deve-se buscar todos os casos que o possuem.
- **Passo 4**: deve-se submeter todos os casos selecionados no **Passo 3** ao processo de classificação, obtendo o número de casos classificados com sucesso.

- **Passo 5**: deve-se alterar a polarização do termo selecionado no **Passo 3** e submeter os casos que o possuem ao processo de classificação.
- **Passo 6**: deve-se verificar a quantidade de casos processados com sucesso nas duas execuções.
    - Se após a modificação de polarização do termo for obtido um número maior ou igual de casos polarizados, deve-se persistir a modificação na base de conhecimento. Caso contrário, deve-se voltar à polarização original e seguir o processo.
- **Passo 7**: efetuar o **Passo 3** até que se tenha verificado todos os termos polarizados.
    - Ao finalizar o **Passo 3**, vá para **Passo 2**.
- **Passo 8**: efetuar o **Passo 2** até que se tenha verificado todos os casos de entrada.
- **Passo 9:** fim do método.

Para atestar a viabilidade do método proposto, ele foi aplicado sobre o domínio de filmes, discutido na seção anterior. Inicialmente, utilizou-se a base de conhecimento original, que contempla o léxico original proposto por Hu e Liu (2004), e a ontologia apresentada na Seção 4.2.2.1. O resultado da execução é apresentado na Tabela 18, a seguir.

Tabela 18 – Medidas do protótipo no domínio de filmes utilizando o método automático de refinamento do léxico

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base$_{manual}$	804	834	192	161	**0,82**	0,81	0,83	0,82
Base$_{automático}$	827	796	169	199	**0,82**	0,83	0,81	0,82

Fonte: Elaborado pelo autor

É possível perceber que se obteve um resultado com a mesma acurácia nas duas execuções, o que demostra a viabilidade da abordagem desenvolvida. Optou-se por submeter novamente os casos que ainda apresentam problemas ao processo de adequação, mais detalhes são apresentados na Tabela 19.

Tabela 19 – Aplicando o método pela segunda vez na base de casos com problema

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base$_{manual}$	804	834	192	161	0,82	0,81	0,83	0,82
Base$_{automático}$	827	796	169	199	0,82	0,83	0,81	0,82
Segunda	842	835	154	160	**0,84**	0,85	0,84	0,84

Fonte: Elaborado pelo autor

Percebe-se que ao submeter a base de casos ao mesmo método mais uma vez, foi possível aumentar a acurácia da classificação, passando de 0,82 para 0,84. O resultado da segunda submissão dos dados está representado pela linha Segunda da Tabela 19.

A Tabela 20 apresenta a aplicação de McNemar entre as execuções apresentadas anteriormente.

Tabela 20 – Significância estatística para teste de viabilidade do método de adequação do léxico

Método 1	**Método 2**	**c_{01}**	**c_{10}**	**McNemar**
$Base_{manual}$	$Base_{automático}$	306	247	6,51
$Base_{automático}$	Segunda	108	54	18,67

Fonte: Elaborado pelo autor

Por meio das acurácias apresentadas na Tabela 19, pode-se verificar que o método automático permite atingir a mesma acurácia que o método manual, o que significa que o método automático pode ser incorporado ao modelo sem problemas. Quando comparados os dois métodos, manual e automático, é possível observar que existe significância estatística, ou seja, o valor de McNemar é superior a 3,841 entre as suas execuções. Ainda pode-se observar que o método automático classificou 306 casos corretamente, os quais o método manual havia errado (c_{01}), e 247 de maneira errada, que o método manual havia acertado (c_{10}). Comparando os dados das colunas c_{01} e c_{10}, percebe-se que é possível ter um ganho (mesmo que não explícito na acurácia) no número de casos classificados corretamente pelo método automático.

Ao executar o método automático novamente, percebeu-se que a acurácia sofreu um aumento no número de acertos e, comparando essa execução com a anterior, obtém-se significância estatística. Como esse método não representa o foco principal da tese, optou-se por não evoluí-lo além dos pontos apresentados nessa execução.

Com o ciclo do modelo proposto quase finalizado, a seção a seguir procura apresentar mais detalhes sobre o desenvolvimento da estratégia de adaptação de um caso passado para a solução de um novo caso.

4.3 CONTRUÇÃO DA ETAPA DE ADAPTAÇÃO DO MODELO

O modelo proposto de tese prevê uma etapa de adaptação, objetivando utilizar o conhecimento adquirido por meio da base de

conhecimentos a partir das classificações anteriores, de modo que se possa aproveitar parte de um caso para auxiliar na polarização de um novo caso.

Analisando os casos que foram submetidos ao processo de polarização do modelo proposto, percebe-se que uma das principais causas de falhas na polarização são os termos ambíguos existentes no léxico de sentimento. A deficiência no tratamento de ambiguidade por parte do modelo também é comum a muitas soluções de análise de sentimento. Segundo Cao, Zhang e Xiong (2015), a tarefa de desambiguação é um grande desafio para a área de análise de sentimento. Infelizmente, muitas pesquisas simplesmente não a consideram em seus modelos, o que na maior parte das vezes torna as suas soluções pouco eficientes.

O modelo proposto neste trabalho possui uma etapa de adequação do léxico baseada em uma base de treinamento ou na própria base de casos já armazenados. Essa etapa auxilia no processo de otimização do léxico para os casos já conhecidos. Para a etapa de adaptação do modelo, optou-se por atacar os termos considerados ambíguos. No contexto da tese, termos ambíguos são aqueles que existem tanto em casos classificados como positivos quanto em casos classificados como negativos.

A etapa de adaptação é iniciada quando um caso não é recuperado da base de casos conhecidos, ou seja, o seu grau de similaridade em relação ao novo caso é inferior ao *threshold*. No momento em que a árvore de sentimento é gerada, os termos com orientação semântica relacionados aos conceitos são verificados. Caso haja termos ambíguos, estes entram para a etapa de adaptação.

As seções seguintes apresentam as estratégias experimentadas juntamente com os resultados obtidos. Para efeito de comparação, utilizou-se a base de filmes a partir da adequação manual do léxico apresentado com mais detalhes na Seção 4.2.3.

4.3.1 Experimento 1 - Neutralização dos termos ambíguos

A primeira estratégia verificada foi a neutralização dos termos ambíguos durante o processo de construção das árvores de sentimento. Para cada termo polarizado (orientação semântica) que tinha relação com algum conceito da árvore gerada, foi verificado se este é ambíguo ou não. Caso o termo seja ambíguo, ele tem o seu grau de polarização (1 ou -1) neutralizado, ou seja, é atribuído o grau 0, não tendo mais impacto no cálculo da polarização. Refere-se a esta estratégia como

Experimento 1 (Exp_1). O resultado obtido é apresentado com mais detalhes na Tabela 21.

Tabela 21 – Medidas a partir da neutralização dos termos ambíguos

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base$_{manual}$	804	834	192	161	0,82	0,81	0,83	0,82
Exp$_1$	802	801	194	194	0,81	0,81	0,81	0,81

Fonte: Elaborado pelo autor

Pode-se perceber que a simples neutralização dos termos ambíguos no processo de polarização de um novo caso não apresenta resultados superiores à execução-base, pois a acurácia ficou menor a partir desse experimento.

Considerando a importância dos termos ambíguos para o processo de polarização, a seção a seguir apresenta uma estratégia em que esses termos são utilizados no processo, mas não necessariamente com a sua classificação original.

4.3.2 Experimento 2 – Verificação da frequência polarizada

A partir da constatação obtida no Experimento 1, em que é possível verificar a importância dos termos ambíguos para o processo de polarização, neste segundo experimento tem-se como foco considerar todos os termos ambíguos, mas não necessariamente utilizando a sua polarização original.

Partindo da árvore de sentimento (conceitos da ontologia relacionados com os termos polarizados) elaborada a partir do conteúdo de um novo caso, são selecionados todos os termos ambíguos que estão presentes na árvore. Para cada termo ambíguo, é verificada a quantidade de casos que o contém, tanto positivos quanto negativos. Tendo em vista os dois conjuntos de casos (positivos e negativos) que possui o termo, é levado em consideração o conjunto que possui mais casos, ou seja, a polarização resultante é a mesma do maior conjunto de casos. Caso os dois conjuntos tenham a mesma quantidade de elementos, mantém-se a polarização original. A Tabela 22 apresenta o resultado obtido a partir desse experimento que está referenciado como Exp_2.

Tabela 22 – Medidas a partir a verificação da frequência polarizada

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base $_{manual}$	804	834	192	161	0,82	0,81	0,83	0,82
Exp$_2$	835	834	161	161	0,84	0,84	0,84	0,84

Fonte: Elaborado pelo autor

Utilizando a adaptação baseada na verificação da frequência polarizada de um termo ambíguo, chega-se a uma acurácia maior do que a execução-base. A acurácia obtida de 0,84 é a mesma obtida pela aplicação da estratégia de adequação automática do léxico quando aplicada sobre a execução-base.

O experimento da seção seguinte tem como objetivo identificar em que cenários os termos ambíguos são positivos ou negativos e utilizar essa informação para a polarização na etapa de adaptação.

4.3.3 Experimento 3 – Polarização contextualizada dos termos ambíguos

O terceiro experimento visa identificar, para cada termo ambíguo, em que contexto ele é utilizado como positivo ou negativo. Para isso, utilizou-se o método de Decomposição de Valores Singulares, do inglês *Singular Value Decomposition* (SVD). Esse método, ao final do processamento, gera uma matriz termo *x* termo, em que na intersecção da linha pela coluna tem-se o número de coocorrências de documentos de uma base.

Gerou-se a matriz SVD de toda a base de casos conhecidos sobre filmes, domínio utilizado para os testes da etapa de adaptação. Para cada orientação semântica, gerou-se uma matriz, ou seja, é possível verificar todos os termos que mais coocorrem com um termo em questão com o seu uso positivo e negativo.

Para cada novo caso é gerado um vetor de termos, no qual se adiciona todos os termos polarizados ou que façam parte da ontologia de domínio. Dentro da análise de cada caso, são verificados todos os termos polarizados, ou seja, se o termo é ambíguo ou não. Caso ele seja ambíguo, são consultadas as duas matrizes SVD (positiva e negativa), de modo a extrair um vetor de termos que coocorrem com o termo ambíguo em questão, no seu uso como positivo e como negativo.

O primeiro experimento executado utilizando SVD é referenciado como Experimento 3 (Exp_3). Nele, é comparado o vetor vindo do novo caso ($vetor_{caso}$) com o vetor de termos positivos ($vetor_{positivo}$) e negativos ($vetor_{negativo}$), utilizando o cálculo de similaridade baseado no cosseno do ângulo entre os vetores. Se o $vetor_{caso}$ for mais próximo do $vetor_{positivo}$, isso indica que o termo ambíguo utilizado no caso é mais similar com o seu uso em outros casos positivos, caso contrário, em casos negativos.

No Experimento 3, leva-se em consideração que, se o termo ambíguo possuir uma similaridade igual ao $vetor_{positivo}$ e ao $vetor_{negativo}$, a

sua polarização é neutralizada. Os resultados obtidos podem ser observados na Tabela 23.

Tabela 23 – Medidas obtidas pelo experimento 3.1

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base $_{manual}$	804	834	192	161	0,82	0,81	0,83	0,82
Exp$_3$	847	842	149	153	0,85	0,85	0,85	0,85

Fonte: Elaborado pelo autor

Percebe-se que a acurácia obtida, 0,85, é superior à acurácia da execução-base, o que demostra que o uso do SVD pode representar uma solução eficiente para a etapa de adaptação modelo proposto.

A seção a seguir apresenta uma análise comparativa dos resultados obtidos pelos experimentos para a tarefa de adaptação. Objetiva-se com esta seção identificar a estratégia mais eficiente para a tarefa de adaptação.

4.3.4 Análise dos resultados obtidos

Os experimentos propostos têm como objetivo definir qual estratégia deve ser adotada para a etapa de adaptação do modelo proposto, atacando os termos ambíguos durante o processo de polarização a partir da árvore de sentimento gerada. A Tabela 24 apresenta todos os resultados obtidos para a definição da estratégia a ser adotada para a etapa de adaptação do modelo de tese.

Tabela 24 – Medidas obtidas a partir dos experimentos para adaptação

Teste$_{filmes}$	**TP**	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Base $_{manual}$	804	834	192	161	0,82	0,81	0,83	0,82
Exp$_1$	802	801	194	194	0,81	0,81	0,81	0,81
Exp$_2$	835	834	161	161	0,84	0,84	0,84	0,84
Exp$_3$	847	842	149	153	**0,85**	0,85	0,85	0,85

Fonte: Elaborado pelo autor

A acurácia máxima entre os experimentos foi de 0,85 obtida no Experimento 3. Para atestar se as abordagens realmente apresentam um ganho no processo de classificação dos casos como positivo ou negativo, as Execuções 1, 2 e 3 foram comparadas com a execução-base, com o intuito de analisar a significância estatística. Mais informações sobre essa comparação são apresentadas na Tabela 25.

Tabela 25 – Significância estatística entre a execução-base com as estratégias candidatas à adaptação

Método 1	**Método 2**	**c_{01}**	**c_{10}**	**McNemar**

Base manual	**Exp_1**	17	52	16,75
Base manual	**Exp_2**	40	9	20,90
Base manual	**Exp_3**	90	39	20,96

Fonte: Elaborado pelo autor

Pode-se observar que o Experimento 1 apresenta um acurácia inferior à execução-base, o que o torna inviável como proposta de adaptação do modelo. Analisando as Execuções 2 e 3, ambas apresentam uma acurácia melhor em relação a execução-base de 0,84 e 0,85, respectivamente. Nesse sentido, pode-se optar pelas duas abordagens como propostas de solução para a adaptação.

Para verificar se a estratégia adotada pelo Experimento 3 possui significância estatística quando comparada ao Experimento 2, calcula-se McNemar dessas duas execuções, a Tabela 26 apresenta mais detalhes sobre essa comparação.

Tabela 26 - Significância estatística entre o Experimento 2 e 3 para a adaptação

Método 1	**Método 2**	**c_{01}**	**c_{10}**	**McNemar**
Exp_2	**Exp_3**	65	45	4,01

Fonte: Elaborado pelo autor

Ao analisar o valor da coluna $c_{01,}$ percebe-se que 65 casos foram classificados de maneira correta pelo Exp3, mas não haviam sido classificados pelo Exp2. Já a coluna c_{10} demostra o contrário, ou seja, 45 casos foram classificados erroneamente pelo Exp3, mas classificados corretamente pelo Exp2. O que se pode perceber dessa análise é que o Exp3 classificou corretamente 20 casos a mais que Exp2, além de possuir significância estatística, como apresentado no valor de McNemar.

Tendo em vista os resultados obtidos sugere-se a utilização de uma matriz SVD como forma de adaptar os termos polares ambíguos a partir de um vetor de contexto.

Vale ressaltar que a solução baseada na frequência polarizada (Experimento 2) não deve ser descartada como uma possível solução de adaptação. O motivo para tal afirmação reside no custo computacional elevado para gerar uma matriz SVD. Nos casos em que não se tenha uma infraestrutura adequada para tal operação, pode-se utilizar a estratégia 2 como solução para a adaptação de termos ambíguos, levando em consideração o contexto da sentença em relação aos demais casos da base.

Tendo uma resposta à tarefa de adaptação definida, a próxima seção apresenta o modelo final da tese a partir de todas as constatações obtidas e análises efetuadas.

4.4 MODELO FINAL

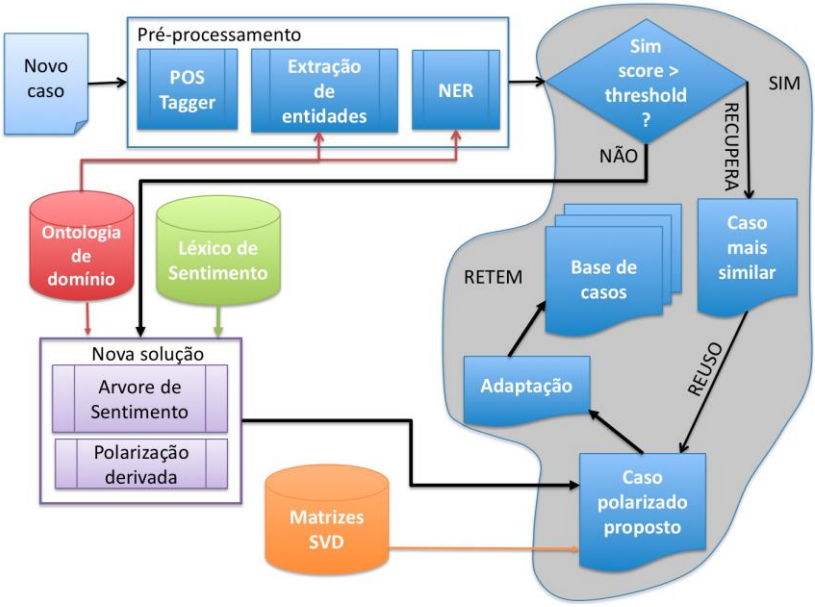
Utilizando como base os experimentos efetuados ao longo desta tese, uma proposta final do modelo foi elaborada. Esta seção tem como objetivo apresentar o modelo final de tese juntamente com uma avaliação quantitativa, análise dos resultados e, por fim, a apresentação de alguns cenários de uso do modelo desenvolvido.

4.4.1 Apresentação do modelo final

O modelo aqui apresentado utiliza como base o modelo desenvolvido e apresentado na Seção 4.1, Figura 29. O motivo para utilizar aquele modelo como base é o sucesso obtido nas avaliações efetuadas.

Existiam alguns pontos que o modelo apresentado na Figura 29 não tratava de maneira adequada. Tendo isso em vista, na Seção 4.2 e 4.3 foram apresentadas algumas estratégias e evoluções para chegar ao modelo final. Um dos pontos principais da evolução do modelo consiste na inclusão da tarefa de adaptação do modelo de tese. Nessa etapa, procura-se tratar os casos que possuem termos polarizados ambíguos, ou seja, que podem apresentar polarização distinta dependendo do cenário em que estão sendo aplicados. A Figura 40 apresenta mais detalhes sobre o modelo resultante.

Figura 40 – Modelo final da tese



Fonte: Elaborado pelo autor

Existem dois elementos novos no modelo apresentado na Figura 40 em relação ao apresentado na Figura 28: (1) o repositório das matrizes SVD e (2) a etapa de adaptação.

A adaptação foi inserida dentro do fluxo RBC (região com o fundo cinza) do modelo, estando entre o Caso polarizado proposto e a persistência do caso na Base de casos. É importante perceber que a etapa de adaptação também tem reflexo sobre a árvore de sentimento gerada para os casos que não foram recuperados, ou seja, que não entraram no ciclo do RBC. A árvore gerada é submetida ao processo de adaptação e todos os termos ambíguos são verificados para constatar se, no contexto, devem sofrer mudanças na sua polarização.

Para entender melhor a etapa de adaptação e o processo completo do modelo proposto, a seção a seguir demonstra o fluxo mais longo e completo do modelo, a fim de ilustrar a sua utilização.

4.4.1.1 Exemplo das etapas do modelo proposto

Para exemplificar as etapas do modelo, utiliza-se como base um *review* de câmera cuja polaridade original é negativa. O conteúdo do *review* é apresentado com mais detalhes a seguir.

Sure my canon sd-30 is super cute and i get lots of compliments on how small and sexy it is, but heaven forbid you want to take pictures with it!! This camera cannot take pictures in low light at all! every picture i've taken at parties has been out of focus, which adds up to a lot of hazy photos of great memories. I took it back to the camera shop and the guy said i had the setting right but "yeah, they don't work well in low light." well if i needed a bright sunny day for every shot i could make a pinhole camera out of a cardboard box!
The pictures it takes in daylight are nice. but anyone who wants a tiny little camera like this is planning on putting it in their pocket and taking it to parties!
I got the sony cybershot 10. 1 and it takes great pictures in low light. later, canon.

A primeira etapa do modelo é a de pré-processamento. Nessa etapa são aplicadas as técnicas de POS *Tagger*, extração e reconhecimento de entidades nomeadas (*named entity recognition* - NER). Para isso, o conteúdo do *review* é dividido em sentenças e cada uma delas terá os seus termos processados, identificando qual a sua classe gramatical, se é um termo polar (possui ou não um grau positivo ou negativo) e se é uma instância ou conceito da ontologia de domínio.

A Figura 41 apresenta um exemplo do processamento realizado sobre a primeira sentença do *review* supracitado.

Figura 41 – Exemplo de uma sentença após a etapa de pré-processamento

```
SENTENÇA => sure my canon sd-30 is super cute and i get lots of compliments on how small
and sexy it is, but heaven forbid you want to take pictures with it!!

PALAVRAS:
     - sure (Adjective) [0]
     - my (Possessive pronoun) [0]
     - canon (Noun, singular or mass) {CONCEPT: camera}
     - sd-30 (null) [0]
     - is (Verb, 3rd person singular present) [0]
     - super (Adjective) [1]
     - cute (Adjective) [1]
     - and (Coordinating conjunction) [0]
     - i (Preposition or subordinating conjuntion) [0]
     - get (Verb, base form) [0]
     - lots (Noun, plural) [0]
     - of (Preposition or subordinating conjuntion) [0]
     - compliments (Noun, plural) [0]
     - on (Preposition or subordinating conjuntion) [0]
     - how (Wh-adverb) [0]
     - small (Adjective) [0]
     - and (Coordinating conjunction) [0]
     - sexy (Adjective) [1]
     - it (Personal pronoun) [0]
     - is, (null) [0]
     - but (Coordinating conjunction) [0]
     - heaven (Noun, singular or mass) [1]
     - forbid (Verb, past tense) [-1]
     - you (Personal pronoun) [0]
     - want (Verb, non-3rd person singular present) [0]
     - to (to) [0]
     - take (Verb, base form) [0]
     - pictures (Noun, plural) {CONCEPT: pictures}
     - with (Preposition or subordinating conjuntion) [0]
     - it!!
```

Fonte: Elaborado pelo autor

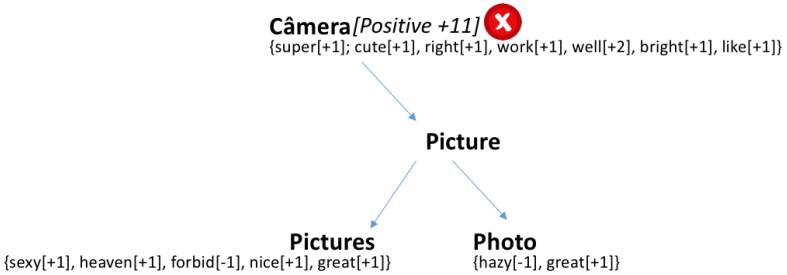
Pode-se observar, no exemplo de saída, os pontos que foram anotados ou polarizados. Como saída da etapa de pré-processamento, têm-se, portanto, os termos com as suas classes (conceitos da ontologia) anotados e os termos polarizados com as suas respectivas polarizações (indicação de 1 ou -1 ao lado de cada termo).

O próximo passo formata o vetor de consulta. Utilizando-se como base o caso apresentado, obtém-se o seguinte vetor: *{canon; super; cute; sexy; heaven; forbid; pictures; camera; picture; hazy; photos; great; camera; right; work; well; bright; nice; like}.*

No caso do domínio de câmeras, o *threshold* adotado foi de 0.43. Esse valor foi obtido por meio do método apresentado na Seção 4.2.2.2. Submetendo o vetor ao processo de recuperação, verifica-se que o caso mais similar possui o seu grau igual a 0.117, ou seja, inferior ao *threshold* utilizado. Com isso o caso recuperado é descartado e o novo caso é submetido para a etapa Nova solução, que tem como objetivo gerar a árvore de sentimento do novo caso e disponibilizá-la para a próxima etapa. A

Figura 42 apresenta a árvore de sentimento inicial desenvolvida na etapa de nova solução.

Figura 42 – Exemplo da árvore gerada na etapa Nova solução

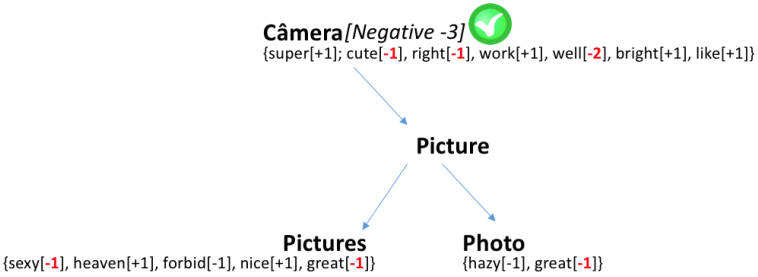


Fonte: Elaborado pelo autor

Como já informado no início desta seção, o *review* utilizado tem polarização original negativa. Por meio da primeira árvore de sentimento gerada apenas utilizando os termos polarizados do léxico, o valor obtido foi +11, ou seja, o caso foi classificado como positivo, o que significa que o caso foi polarizado errado. Antes de desenvolver a etapa de adaptação, esse seria o resultado disponível para o usuário e para armazenado na base de casos.

O caso polarizado proposto, que possui o conteúdo do novo caso, somado à árvore de sentimento gerada, é submetido à etapa de adaptação. Neste ponto, todos os termos caracterizados como ambíguos, utilizados tanto por casos positivos quanto por casos negativos, são analisados para verificar se mantêm a polarização original ou devem mudar de polarização (maiores detalhes são descritos na Seção 4.3.3). A Figura 43 apresenta a árvore adaptada a partir da análise dos termos ambíguos utilizando as matrizes SVD.

Figura 43 – Exemplo de árvore adaptada



Fonte: Elaborado pelo autor

Como pode ser observado na Figura 43, vários termos polares tiveram a sua polarização alterada pelo processo de adaptação. Todos os termos que tiveram a sua polarização modificada foram destacados em vermelho. A partir do processamento da árvore de sentimento original pela etapa de adaptação, obtém-se uma nova polarização do caso como negativa, que é a polarização de fato do novo caso. Esse caso polarizado proposto é, então, armazenado na base de casos.

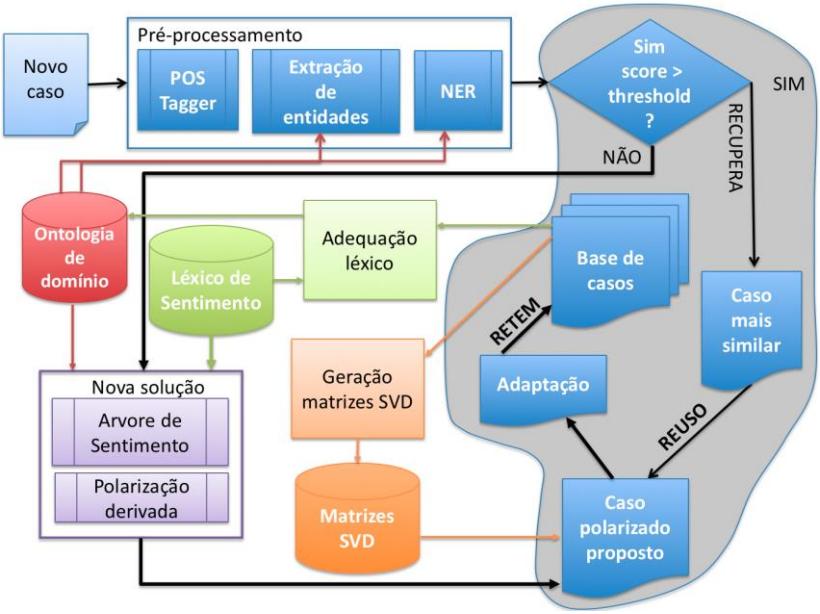
Para promover suporte ao modelo proposto, foram desenvolvidos alguns métodos e implementações. Na próxima seção são apresentados maiores detalhes.

4.4.1.2 Aplicações de suporte ao modelo

A partir de desenvolvimento e da implementação do modelo de tese, identificaram-se alguns pontos que poderiam ser automatizados na forma de aplicações. Tendo isso em vista, desenvolveram-se algumas ferramentas que podem contribuir com o funcionamento do modelo. Vale lembrar que esses itens não entram no modelo de tese, pois apenas dão apoio a etapas específicas, podendo ser facilmente substituídos ou até mesmo desconsiderados.

A Figura 44 apresenta as etapas, não obrigatórias, bem como a forma como elas interagem com as outras etapas específicas do modelo de tese.

Figura 44 – Aplicações de suporte ao modelo de tese



Fonte: Elaborado pelo autor

Pode-se observar na Figura 44, que foram adicionados ao modelo dois novos processos (etapas). O primeiro chamado de Adequação léxico e o outro chamado de Geração matrizes SVD.

A etapa Adequação léxico tem como objetivo adaptar o léxico original proposto pode Hu e Liu (2004) para o domínio de aplicação específico. Essa etapa considera os termos polares no léxico de sentimento e verifica a sua utilização em relação à base de casos já polarizada. Se os termos ocorrerem mais vezes em casos positivos, eles são transformados em positivos, caso contrário, são transformados em negativos. Todas essas modificações são persistidas na ontologia de domínio, podendo depois ser facilmente utilizadas no processo de polarização durante a construção da árvore de sentimento. Mais informações sobre a etapa de adequação do léxico estão declaradas na Seção 4.2.5.

A etapa denominada de Geração matrizes SVD tem como objetivo gerar as matrizes SVD para cada domínio de aplicação, utilizando como fonte a base de casos. A Seção 2.4.3.8 apresenta mais informações sobre o processo de geração das matrizes baseadas em SVD. De maneira geral, para cada termo polar ambíguo, é verificado o conjunto de palavras que mais coocorrem dentro de uma polarização

específica (conjunto de casos positivos ou negativos), ou seja, é elaborado um vetor de termos para cada classificação (positiva/negativa). Na sequência, a etapa de adaptação precisa apenas solicitar o vetor positivo e negativo do termo polar no determinado domínio e compará-lo ao vetor do caso polarizado proposto, podendo definir se o termo é positivo ou negativo dependendo da proximidade dos vetores.

Estando o modelo final definido conjuntamente aos processos e sistemas que o orbitam, parte-se para a etapa de avaliação do protótipo desenvolvido a partir do modelo. A seção seguinte tem como objetivo apresentar mais detalhes sobre o processo de avaliação do modelo desenvolvido.

4.4.2 Avaliação do modelo

O modelo de tese foi implementado na forma de um protótipo funcional. O protótipo respeita todas as etapas e especificações propostas no modelo. Para a avaliação do modelo desenvolvido, utilizaram-se as duas bases de dados já apresentadas na tese: *reviews* de câmeras e de filmes. Nesta seção, tem-se como objetivo avaliar o modelo de tese a partir da sua implementação. Para isso, procura-se isolar algumas partes para mensurar a contribuição no modelo final.

Entende-se que a etapa de geração das árvores de sentimento é permitida a partir dos conceitos da ontologia de domínio combinados com a polaridade vinda de um léxico. Como o núcleo do modelo, a árvore de sentimento não pode ser excluída de qualquer execução.

Os pontos que são avaliados a partir da execução-base utilizando somente a geração da árvore de sentimento com a ontologia e o léxico original são: o uso do RBC e o tratamento de negação. O Quadro 5 apresenta mais detalhes sobre as execuções.

Quadro 5 – Nome atribuído para as execuções de avaliação

	RBC?	**Negação?**	**Domínio**
Execução$_1$	Não	Não	Filme
Execução$_2$	Não	Não	Câmera
Execução$_3$	Sim	Não	Filme
Execução$_4$	Sim	Não	Câmera
Execução$_5$	Não	Sim	Filme
Execução$_6$	Não	Sim	Câmera
Execução$_7$	Sim	Sim	Filme
Execução$_8$	Sim	Sim	Câmera

Fonte: Elaborado pelo autor

Pode-se observar que as Execuções 1 e 2 representam, respetivamente, a execução-base para filme e para câmera. As Execuções 3 e 4 apresentam o resultado utilizando o modelo de tese mais o uso do RBC, enquanto que as Execuções 5 e 6 apresentam a contribuição do tratamento da negação considerando a execução-base (1 e 2). Por fim, as Execuções 7 e 8 representam a combinação da execução-base com o RBC e o tratamento de negação. O resultado para as execuções foram os apresentados na Tabela 27.

Tabela 27 – Execuções de avaliação

	TP	**TN**	**FP**	**FN**	**Acu.**	**Prec.**	**Rec.**	**F**
Execução$_1$	804	667	192	328	0,739	0,807	0,710	0,756
Execução$_2$	664	385	62	332	0,727	0,915	0,667	0,771
Execução$_3$	771	671	225	324	0,724	0,774	0,704	0,737
Execução$_4$	674	488	52	229	0,805	0,928	0,746	0,828
Execução$_5$	800	693	196	302	0,750	0,803	0,726	0,763
Execução$_6$	663	481	63	236	0,793	0,913	0,737	0,816
Execução$_7$	768	684	228	311	0,729	0,771	0,712	0,740
Execução$_8$	673	555	53	162	0,851	0,927	0,806	0,862

Fonte: Elaborado pelo autor

As próximas execuções procuram identificar o impacto da inserção do processo de adequação do léxico e da adaptação a partir da execução-base. O Quadro 1 apresenta os nomes das próximas execuções.

Quadro 6 – Nomes das execuções de avaliação utilizando adequação do léxico e adaptação

	Adequação Léxico?	**Adaptação?**	**Domínio**
Execução$_9$	Sim	Não	Filme
Execução$_{10}$	Sim	Não	Câmera
Execução$_{11}$	Não	Sim	Filme
Execução$_{12}$	Não	Sim	Câmera

Fonte: Elaborado pelo autor

As Execuções 9 e 10 apresentam o uso do processo, não obrigatório, de adequação do léxico. Esse processo consiste na adequação dos termos polares do léxico para um domínio específico utilizando como base as Execuções-base 1 e 2. Nas Execuções 11 e 12,

procura-se verificar o benefício da etapa de adaptação a partir das Execuções-base 1 e 2. A Tabela 28 apresenta os resultados obtidos.

Tabela 28 – Execuções de avaliação utilizando adequação e adaptação

	TP	TN	FP	FN	Acu.	Prec.	Rec.	F
Execução$_9$	834	847	162	148	0,844	0,837	0,849	0,843
Execução$_{10}$	627	601	99	116	0,854	0,864	0,844	0,854
Execução$_{11}$	894	789	102	206	0,845	0,898	0,813	0,853
Execução$_{12}$	676	549	50	168	0,849	0,931	0,801	0,861

Fonte: Elaborado pelo autor

O último conjunto de execuções, 13 e 14, representam, respectivamente, a solução completa para o contexto de filme e de câmera. Entende-se como solução completa aquela que possui a execução-base, utiliza RBC, tratamento de negação, adequação do léxico e adaptação. A Tabela 29 apresenta os resultados obtidos.

Tabela 29 – Execuções de avaliação utilizando o modelo completo

	TP	TN	FP	FN	Acu.	Prec.	Rec.	F
Execução$_{13}$	847	842	149	153	0,850	0,850	0,847	0,849
Execução$_{14}$	653	660	73	57	0,910	0,899	0,920	0,909

Fonte: Elaborado pelo autor

Tendo as execuções definidas e os resultados obtidos, a próxima seção tem como objetivo analisar os dados gerados, possibilitando a avaliação do modelo de tese.

4.4.3 Análise dos resultados

A seção anterior, 4.3.2, apresentou uma série de execuções para que fosse possível avaliar cada elemento que compõe o modelo proposto. Esta seção, por sua vez, tem como objetivo analisar os resultados obtidos pelas execuções anteriores.

Para facilitar a análise, divide-se a acurácia dos resultados das execuções em subgrupos positivo e negativo. A Tabela 30 apresenta esses valores.

Tabela 30 – Subgrupo da acurácia obtida pelas execuções

	Acurácia Positivos	Acurácia Negativos	Acurácia Total
Execução$_1$	0,807	0,670	**0,739**
Execução$_2$	0,915	0,537	**0,727**
Execução$_3$	0,774	0,674	**0,724**
Execução$_4$	0,928	0,681	**0,805**
Execução$_5$	0,803	0,696	**0,749**
Execução$_6$	0,913	0,670	**0,792**

Execução$_7$	0,771	0,687	**0,729**
Execução$_8$	0,926	0,774	**0,851**
Execução$_9$	0,837	0,851	**0,844**
Execução$_{10}$	0,864	0,838	**0,854**
Execução$_{11}$	0,898	0,793	**0,845**
Execução$_{12}$	0,931	0,766	**0,849**
Execução$_{13}$	0,850	0,846	**0,850**
Execução$_{14}$	0,900	0,920	**0,910**

Fonte: Elaborado pelo autor

A partir dos dados apresentados nas Tabela 30 e Tabela 31, pode-se reparar que a inclusão do RBC ao modelo – Execuções 3 e 4 – aumentou a acurácia para o domínio de câmera – Execução 2 e 4 – de 0,727 para 0,805. Esse aumento também ocorre para os subgrupos positivo e negativo do domínio, principalmente nos casos negativos, de 0,537 para 0,681.

No caso do domínio de filmes – Execuções 1 e 3 – a acurácia total foi diminuída de 0,739 para 0,724 quando adicionado o RBC. Quando se observa apenas o subgrupo de casos negativos, a acurácia foi melhorada de 0,670 para 0,674. O problema está nos casos positivos que tiveram uma diminuição da sua acurácia de 0,807 para 0,774. A Tabela 31 apresenta o valor de McNemar das execuções do domínio de filme (1 e 3) e câmera (2 e 4).

Tabela 31 – McNemar da execução base com RBC

Método 1	**Método 2**	**c_{01}**	**c_{10}**	**McNemar**
Execução$_1$	**Execução$_3$**	75	104	4,380
Execução$_2$	**Execução$_4$**	124	11	96,267

Fonte: Elaborado pelo autor

O motivo para tal divergência entre a aplicação do RBC no domínio de câmera e no o domínio de filmes está relacionado ao método adotado para calcular o *threshold.* Como esse método utiliza apenas a média de todos os casos similares da base (ver Seção 4.2.2.2) por domínio, em casos de bases com assuntos mais subjetivos, como a classificação de um filme, deve-se adotar um *threshold* mais alto para não utilizar casos não tão similares. Mais uma alternativa para esse ponto, seria buscar outra equação ou até mesmo outra estratégia para definir o *threshold*.

Percebe-se que, em ambas as situações, existe significância estatística entre as execuções, o que torna a diferença das execuções relevantes.

O segundo grupo de execuções analisa a inclusão do tratamento de negação, descrito na Seção 4.1.3.1. Em resumo, quando se encontra um termo de negação em uma sentença, são analisados os quatro termos anteriores e os quatro termos posteriores e, caso exista alguma palavra polar (que tenha conotação positiva ou negativa segundo o léxico) nesse intervalo, a polarização do termo é alterada.

Ao utilizar o tratamento de negação em ambos os cenários, câmera – Execução 6 – e filmes – Execução 5 –, obteve-se um aumento na acurácia total. Os dois cenários apresentam uma leve diminuição na acurácia dos casos positivos, câmera de 0,915 para 0,913 e filmes de 0,807 para 0,803, conforme apresentado na Tabela 30. Contudo, quando se analisa os casos negativos de maneira isolada, percebe-se um grande aumento das acurácias, para câmera de 0,537 para 0,670 – Execuções 2 e 6, respectivamente – e para filmes de 0,670 para 0,696 – Execuções 1 e 5. O motivo para o aumento da classificação correta dos casos negativos pode ser explicado pela utilização de termos positivos em *reviews* que possuem sentido negativo, como, por exemplo: "Esse filme não é bom!". O que se percebe é que essa é uma prática comum pelos consumidores que escrevem *reviews* sobre produtos, já que em ambos os domínios teve um expressivo aumento na acurácia. A Tabela 32 apresenta o valor de McNemar entre as execuções.

Tabela 32 – McNemar da execução-base com tratamento de negação

Método 1	**Método 2**	c_{01}	c_{10}	**McNemar**
Execução$_1$	**Execução$_5$**	87	65	3,480
Execução$_2$	**Execução$_6$**	135	40	52,663

Fonte: Elaborado pelo autor

Percebe-se que, estatisticamente, o aumento obtido com o uso do tratamento de negação no domínio de filme não obteve significância estatística, pois o valor de McNemar ficou abaixo de 3.841, ou seja, as execuções aconteceram de maneira muito similar. Analisando o domínio de câmera, existe significância estatística. Ao avaliar a coluna c_{01} percebe-se que foram classificados 135 casos corretamente pela Execução 6 que não haviam sido classificados pela Execução 1. A coluna c_{10} ilustra que a Execução 6 classificou 40 casos de maneira errada, os quais antes eram classificados corretamente pela Execução 2. O que se pode perceber é que, proporcionalmente, a Execução 6 apresenta muito mais benefícios que malefícios em relação à Execução2.

O terceiro grupo de execução tem como objetivo combinar o uso do RBC com o tratamento de negação às execuções-base dos domínios

de câmera e filmes. Percebe-se que o uso combinado do RBC com o tratamento de negação teve resultados bem distintos nos domínios de aplicação. Para o domínio de câmera, a acurácia foi elevada consideravelmente em relação à execução-base e demais execuções, passando de 0,727 para 0,851. Nesse caso, tanto o uso da negação quando o uso de RBC contribuiu para o aumento total da acurácia. Quando aplicado ao domínio de filmes, a acurácia total teve uma queda de 0,739 para 0,729. Ao analisar os resultados por subgrupos positivos e negativos, a acurácia dos casos negativos sofreu um aumento de 0,670 para 0,687, enquanto que no subgrupo positivo, percebe-se uma queda na acurácia, de 0,807 para 0,771, influenciando diretamente na queda da acurácia total.

O motivo para a queda da acurácia total está no cálculo do *threshold*, fazendo com que casos que não são realmente similares sejam entregues como proposta de solução. Isso reforça a importância de se buscar como trabalho futuro um método mais eficiente para calcular o *threshold*.

O ganho apresentado no domínio de câmeras é muito maior que a queda da acurácia no domínio de filmes. Por conta desse fato, se manteve tanto o RBC como o tratamento de negação no modelo final da tese. A Tabela 33 detalha a significância estatística entre as execuções.

Tabela 33 – McNemar da execução-base com RBC e tratamento de negação

Método 1	**Método 2**	**c_{01}**	**c_{10}**	**McNemar**
Execução$_1$	**Execução$_7$**	141	160	1,076
Execução$_2$	**Execução$_8$**	215	36	129,084

Fonte: Elaborado pelo autor

Pode-se verificar que no caso do domínio de filmes, Execuções 1 e 7, a diferença de execução não possui significância estatística, valor de McNemar é inferior a 3,841. Ou seja, as execuções são muito similares e não apresentam um ganho ou perda significante. Quando se analisa o domínio de câmera, Execuções 2 e 8, percebe-se uma melhoria bastante significativa. Ao analisar a coluna c_{01} da Tabela 33, pode-se perceber que a Execução 8 classificou 215 casos corretamente antes não classificados pela Execução 2. A coluna c_{10} ilustra 36 casos que antes eram classificados de maneira correta pela Execução 2 e foram classificados errados pela Execução 8, o que demostra a grande evolução dos acertos por parte da Execução 8.

O quarto grupo de execuções avaliou a inserção da etapa de adequação do léxico, Execuções 9 e 10, que consiste na identificação

dos termos que devem ser desconsiderados ou que devem ter a sua polarização alterada segundo um conjunto de casos de um domínio frente às execuções-base, 1 e 2. Em ambos os domínios, a acurácia foi elevada em relação às execuções-base. No domínio de câmera, Execuções 2 e 10, a acurácia foi de 0,727 para 0,851, já no domínio de filmes, Execuções 1 e 9, a acurácia total foi de 0,739 para 0,844. Pode-se observar que, ao utilizar a etapa de adequação do léxico na execução-base, o número de casos polarizados corretamente nos subgrupos positivo e negativo de um domínio fica bastante próximo, ou seja, esse método ajuda a balancear os acertos entre casos positivos e negativos, conforme pode-se observar na Tabela 30.

A Tabela 34 apresenta mais detalhes sobre a aplicação do McNemar para as execuções-base, Execução 1 e 2, e para as execuções que utilizam adequação de léxico, Execuções 9 e 10.

Tabela 34 – McNemar da execução-base com adequação do léxico por domínio

Método 1	**Método 2**	c_{01}	c_{10}	**McNemar**
Execução$_1$	**Execução$_9$**	295	85	117,161
Execução$_2$	**Execução$_{10}$**	267	88	91,268

Fonte: Elaborado pelo autor

Todas as execuções utilizadas no quarto grupo apresentam significância estatística, ou seja, a adição da etapa de adequação do léxico – Execuções 9 e 10 – não executam igualmente às execuções-base – Execuções 1 e 2 –, mas foram realizadas nas mesmas condições e a melhoria na acurácia não foi gerada ao acaso. Ao analisar as colunas c_{01} e c_{10} pode-se perceber que houve um grande ganho ao adicionar essa etapa ao modelo como um todo.

O quinto grupo de execuções, 11 e 12, tem como objetivo avaliar a inclusão da etapa de adaptação utilizando vetores extraídos das matrizes SVD. Esse processo apresenta um aumento na acurácia total nos domínios utilizados. Para câmera, o valor foi incrementado de 0,727 para 0,849 – Execuções 2 e 12 –, enquanto que no caso do domínio de filmes, o valor foi incrementado de 0,739 para 0,845 – Execuções 1 e 11. O mesmo aumento ocorre para os subgrupos positivo e negativo, apresentados na Tabela 30.

No caso da utilização da adaptação utilizando SVD, não se obtém uma polarização tão balanceada entre casos positivos e negativos. Pode-se perceber que, em ambos os domínios, houve mais acertos para os casos positivos. A Tabela 35 apresenta o valor de McNemar, o qual

atesta a relevância estatística dos valores encontrados entre as execuções.

Tabela 35 – McNemar da execução-base com a etapa de adaptação

Método 1	**Método 2**	**c_{01}**	**c_{10}**	**McNemar**
Execução$_1$	**Execução$_{11}$**	249	37	158,633
Execução$_2$	**Execução$_{12}$**	213	37	125,316

Fonte: Elaborado pelo autor

Pode-se observar na Tabela 35, que a adição da etapa de adaptação, Execuções 11 e 12, ao modelo-base, Execuções 1 e 2, obtiveram significância estatística. Ao analisar a coluna c_{01} e a coluna c_{10} pode-se constatar que o ganho na classificação final por parte do uso da adaptação trouxe muito mais benefício (c_{01}) que malefício (c_{10}) ao processo de classificação como um todo.

No sexto e último grupo de execuções, resolveu-se combinar todos os elementos apresentados no caso-base, ou seja, além da geração das árvores de sentimento, utiliza-se RBC, tratamento de negação, léxico já adaptado para os domínios em questão e adaptação utilizando SVD. Como pode-se observar na Tabela 30, os resultados apresentados pelo modelo final e completo, Execuções 13 e 14, apresentam as melhores acurácias. No domínio de filmes, Execução 13, a acurácia foi de 0,850, enquanto que para o domínio de câmera, Execução 14, a acurácia final atingiu 0,910. Esses foram os melhores resultados obtidos, o que demonstra que a orquestração do modelo completo pode beneficiar-se com cada uma das etapas e métodos utilizados. A Tabela 36 apresenta a significância estatística entre as execuções.

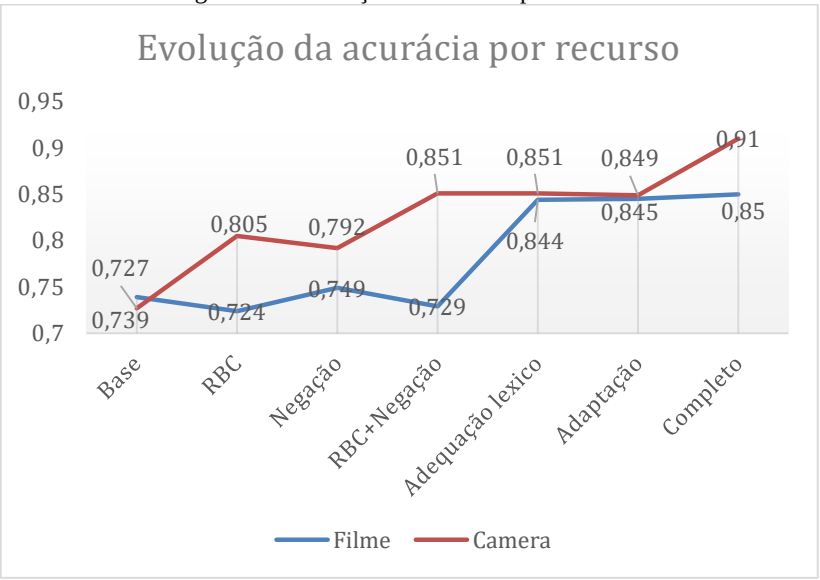
Tabela 36 – McNemar da execução-base com o modelo completo

Método 1	**Método 2**	**c_{01}**	**c_{10}**	**McNemar**
Execução$_1$	**Execução$_{13}$**	376	158	89,815
Execução$_2$	**Execução$_{14}$**	332	68	175,563

Fonte: Elaborado pelo autor

Como é possível observar nos dados da Tabela 36, os resultados obtidos no último grupo de execução possuem significância estatística. Para facilitar a análise da evolução da acurácia total a partir da inclusão dos recursos apresentados, formulou-se a Figura 45.

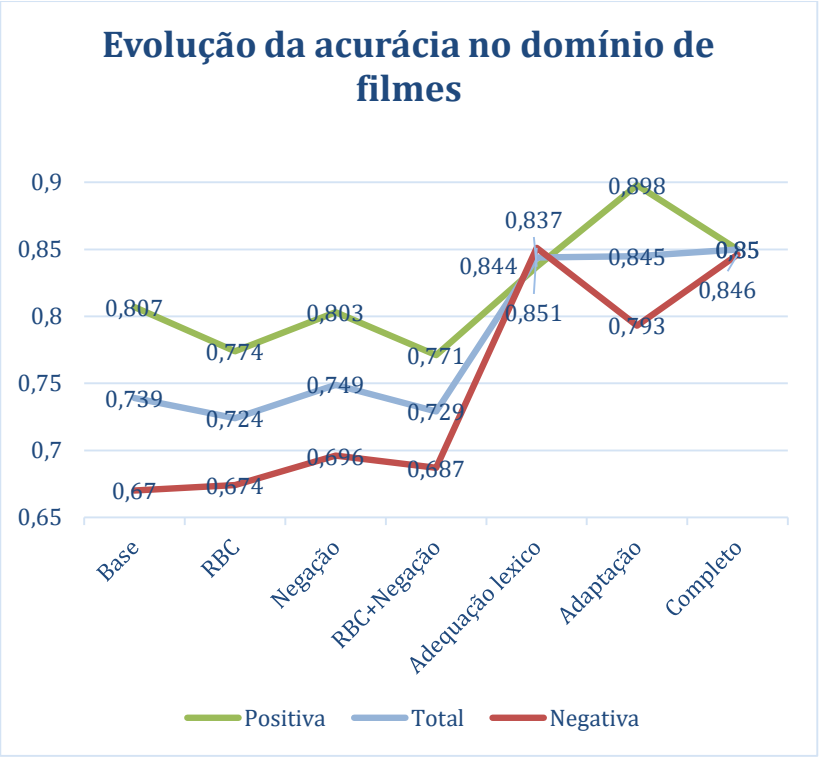
Figura 45 – Evolução da acurácia por recursos



Fonte: Elaborado pelo autor

A Figura 45 ilustra como cada elemento contribui para a evolução do modelo-base até chegar ao modelo final da tese. As Figura 46 e Figura 47 apresentam essa evolução por domínio, levando em consideração os subgrupos positivo/negativo.

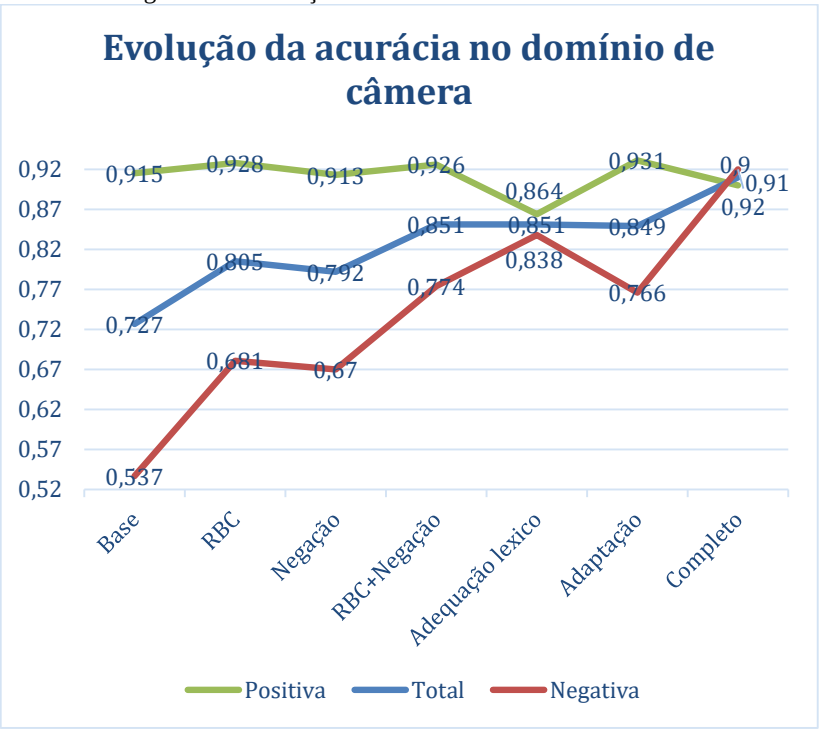
Figura 46 – Evolução da acurácia no domínio de filmes



Fonte: Elaborado pelo autor

A Figura 46 apresenta a evolução das acurácias totais, dos casos positivos e dos casos negativos pelos elementos do modelo de tese. Pode-se perceber que o tratamento dos casos negativos é sempre mais complicado que os dos casos positivos no contexto do domínio dos filmes. Percebe-se que o tratamento de negação auxilia no aumento do acerto das classificações dos casos negativos, mas é a adequação do léxico que garante um balanceamento mais adequado da polarização entre casos positivos e negativos, chegando a melhor acurácia total com o modelo completo.

Figura 47 – Evolução da acurácia no domínio de câmera



Fonte: Elaborado pelo autor

A Figura 47 apresenta a evolução da acurácia no domínio de câmera. Nessa ilustração, é possível perceber que a mesma tendência apresentada na Figura 46 da aplicação do modelo no domínio de filme, também acontece no domínio de câmera, o que demonstra ser uma tendência da aplicação do modelo.

Analisando as Figura 46 e Figura 47, é possível perceber que o comportando da aplicação do modelo em ambos os domínios é parecido, o que demonstra a importância de cada elemento para atingir a acurácia final.

De maneira geral, a presente seção até esse ponto se concentrou na análise do modelo da tese, levando em consideração as suas características e os recursos a partir do resultado da acurácia de acerto nas classificações. Para atestar a relevância do modelo em relação a outras soluções, optou-se por comparar os resultados obtidos com os das duas técnicas que estão entre as mais utilizadas para a área de análise de sentimento. Isso pode ser observado na Tabela 2, que apresenta os

resultados da aplicação do modelo da tese e das técnicas Naïve Bayes (NB) e *Suport Vector Machine* (SVM), usando os mesmos dados. A Tabela 37 apresenta os resultados obtidos.

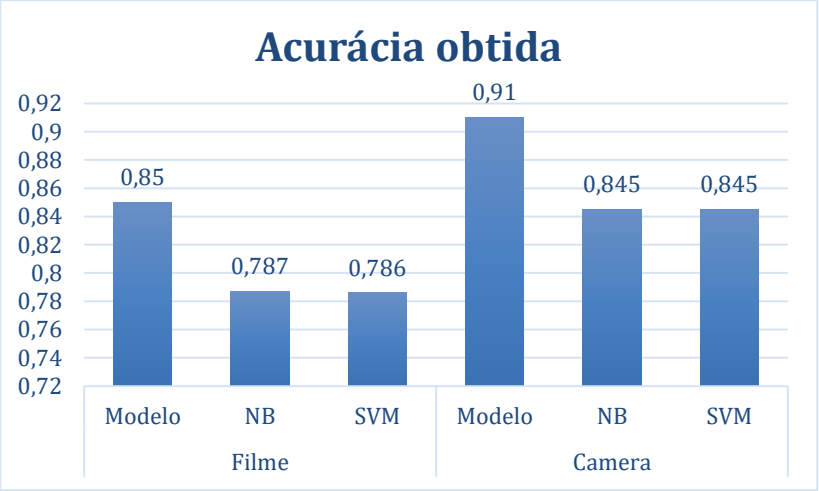
Tabela 37 – Resultados do modelo, NB e SVM

	TP	TN	FP	FN	Acu.	Prec.	Rec.	F
Modelo$_{FILME}$	847	842	149	153	**0,850**	0,850	0,847	0,849
NB$_{FILME}$	822	744	174	251	0,787	0,766	0,825	0,795
SVM$_{FILME}$	785	779	211	216	0,786	0,784	0,788	0,786
Modelo$_{CAMERA}$	653	660	73	57	**0,910**	0,899	0,920	0,909
NB$_{CAMERA}$	620	600	106	117	0,845	0,854	0,841	0,848
SVM$_{CAMERA}$	595	625	131	92	0,845	0,820	0,866	0,842

Fonte: Elaborado pelo autor

Pode-se observar na Tabela 37 que, tanto para o domínio de filmes como para o de câmera, a acurácia total do modelo foi superior a das demais abordagens. A Figura 49 apresenta um gráfico de comparação das acurácias obtidas.

Figura 48 – Comparação da acurácia do modelo com NB e SVM



Fonte: Elaborado pelo autor

O gráfico apresentado na Figura 48 demostra como nos dois domínios a acurácia do modelo foi superior a das demais abordagens. Para atestar se as diferenças entre as acurácias possuem relevância estatística, na Tabela 38 são apresentados os valores de McNemar.

Tabela 38 – McNemar da execução do modelo com NB e SVM

Execução 1	Execução 2	c_{01}	c_{10}	Domínio	McNemar
Modelo	**NB**	203	321	Filme	26,124
Modelo	**SVM**	200	325	Filme	29,288
Modelo	**NB**	83	160	Câmera	23,769
Modelo	**SVM**	87	179	Câmera	31,131

Fonte: Elaborado pelo autor

Pode-se perceber que todas as comparações dos resultados das execuções produzem significância estatística, ou seja, os métodos comparados não oferecem resultados de maneira igual. Analisando os dados das colunas c_{01} e c_{10} pode-se chegar a uma análise mais precisa.

Focando inicialmente no domínio de filmes, pode-se observar que os dados apresentados na coluna c_{01} e c_{10} são muito próximos. Quando se compara o modelo de tese com o NB, pode-se observar que 321 casos são classificados de maneira correta pelo modelo e não pelo NB, analisando a outra coluna, percebe-se que 203 casos foram classificados corretamente pelo NB e classificados errôneamente pelo modelo. Pode-se perceber que além do aumento na acurácia, o modelo apresenta uma melhoria na classificação em relação à aplicação do NB.

Esse mesmo comportamento acontece quando se compara o modelo de tese com NB no domínio de câmera. Também se obtém essa mesma situação quando se compara o modelo com NB e SVM no domínio de filmes.

Tendo comprovado a maior eficiência de classificação do modelo em relação às técnicas de NB e SVM, a próxima seção tem como objetivo apresentar os possíveis cenários de aplicação para o modelo de tese.

4.4.4 Cenários de aplicação do modelo da tese

O modelo da tese foi voltado à classificação com foco na análise de sentimento. Todos os cenários descritos no referencial teórico, mais especificamente na Seção 2.4.4, apresentam situações em que a análise de sentimento pode contribuir, como, por exemplo:

- Análise de dados financeiros;
- Uso na política;
- Análise de *reviews*;
- Análise de trabalhos científicos;
- Detecção de crimes e de terrorismo;
- *Marketing*; e

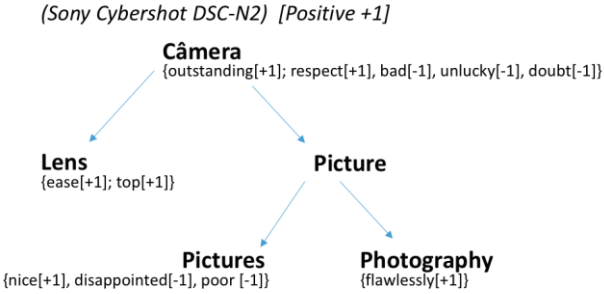
- Sistemas de recomendação.

A proposição desta tese pode contribuir para todos os exemplos apresentados, como qualquer outra solução visando a análise de sentimento. Todavia, existe uma contribuição específica sendo representada pela base de casos gerada.

A base de casos armazena todo o conteúdo original do caso em questão, acrescida da sua árvore de sentimento. Essa árvore apresenta os conceitos e as instâncias da ontologia de domínio bem como os termos polares relacionados, o que permite explicar qual foi o caminho para chegar à polarização final. Neste sentido, a base pode ser utilizada como fonte de informações para sistemas baseados em conhecimento.

A partir da base de casos, pode-se analisar um elemento específico do modelo perante todos os casos já processados e entender o motivo para tal polarização. Por exemplo, caso deseje saber mais informações sobre a câmera Sony Cybershot DSC-N2, a partir da análise dos casos armazenados pode-se chegar a resposta representada pela Figura 49.

Figura 49 – Árvore de sentimento gerada para a câmera DSC-N2



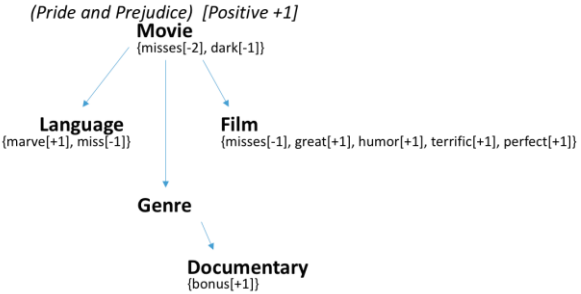
Fonte: Elaborado pelo autor

É possível observar que, a partir de todos os casos armazenados na base de casos sobre a câmera Sony Cybershot DSC-N2, a imagem dela apresenta-se como positiva. Pode-se entender o porquê dessa classificação a partir da árvore de sentimento gerada, além de ver como os conceitos relacionados também estão classificados.

Essas informações podem ser utilizadas como subsídio para inferências de sistemas baseados em conhecimento, ainda mais por ter o conhecimento de que os elementos das árvores de sentimento estão ligados diretamente a elementos da ontologia de domínio, o que permite utilizar e reutilizar tais informações para o apoio à decisão.

Pode-se verificar o mesmo comportamento em outros domínios de aplicação, como, por exemplo, quando se busca a polaridade de alguns filmes, em especial utilizando como base todos os casos já classificados. A Figura 50 apresenta a árvore gerada para o filme Orgulho e Preconceito (*Pride and Prejudice*).

Figura 50 – Árvore de sentimento gerada para o filme Orgulho e preconceito.



Fonte: Elaborado pelo autor

É possível analisar o filme pela perspectiva de outras pessoas e verificar características ou componentes que essas pessoas gostaram ou não do filme. A árvore de sentimento armazenada pode ser utilizada para outras análises ou para a criação de novas árvores. Esse recurso é uma contribuição adicional do modelo como proposta de solução para problemas de classificação, mais especificamente na análise de sentimento.

A próxima seção tem como objetivo comparar o modelo de tese com outros modelos atuais, cujo focoo de atuação seja o mesmo, utilizando como base os critérios levantados na problemática.

4.5 ANÁLISE COMPARATIVA ENTRE MODELOS SIMILARES

O modelo proposto não se assemelha aos modelos encontrados durante o processo de revisão sistemática, os quais foram apresentados na Seção 2.1. Contudo, em função da data de conclusão da revisão sistemática até a data de finalização da concepção do modelo, optou-se pela atualização do estado da arte.

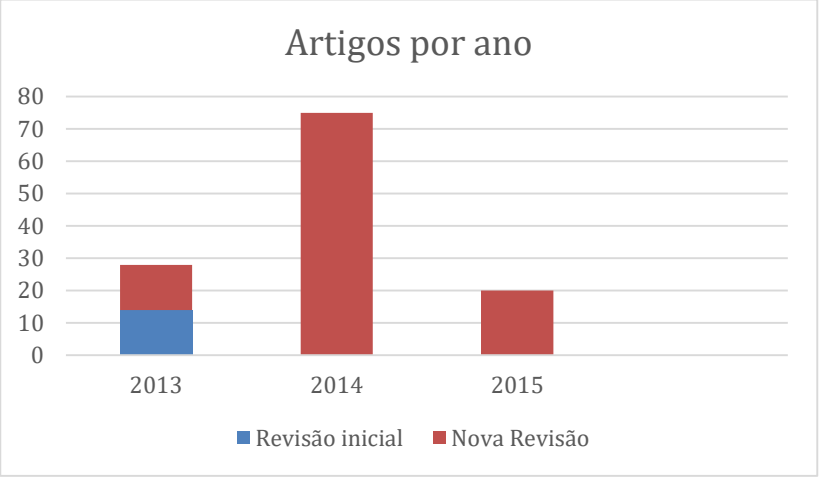
A Seção 4.5.1 apresenta a atualização do estado da arte, buscando apresentar e analisar os novos trabalhos publicados na área, bem como identificar se, no período de desenvolvimento do trabalho, surgiram modelos similares ao proposto nesta tese. A seção seguinte, 4.5.2, tem, por sua vez, como objetivo apresentar uma análise comparativa entre o modelo de tese e os similares encontrados.

4.5.1 Atualização do estado da arte

Para este processo de análise do estado da arte, utilizou-se como base o mesmo protocolo utilizado e descrito na Seção 2.1. A diferença reside no filtro de ano de início das publicações, que neste caso foi 2013 (último ano analisado na primeira revisão sistemática).

A partir da busca efetuada para artigos presentes da base *Web of Science* a partir do ano de 2013, foram encontrados 572 artigos publicados. Dos 572 artigos, estavam relacionados com o tema e disponíveis para *download*, apenas 109. A Figura 51 apresenta mais detalhes sobre a distribuição dos artigos por ano.

Figura 51 – Distribuição dos artigos da área a partir de 2013



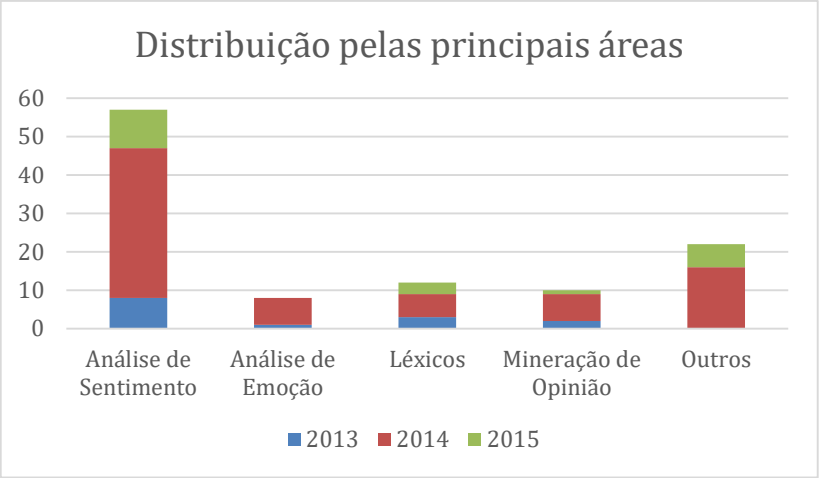
Fonte: Elaborado pelo autor

Para o ano de 2013 foram coletados 14 novos artigos que não haviam sido publicados até a data da revisão inicial, totalizando 28 artigos coletados para esse ano (para esta seção, a partir deste ponto, considerar-se-á apenas os 14 novos artigos coletados para 2013), enquanto que no ano de 2014 foram coletados 75 artigos e, em 2015, 20 artigos.

É importante lembrar que haviam 572 artigos disponíveis para esse intervalo de anos, só não se trabalhou com todos eles pela indisponibilidade de *download*. Esse número demonstra que a área da tese continua em evidência.

A partir da distribuição dos artigos coletados por ano, analisou-se o foco dos artigos frente às subáreas ou às áreas relacionadas com a análise de sentimento. A Figura 52 apresenta mais detalhes.

Figura 52 – Distribuição dos artigos por área temática



Fonte: Elaborado pelo autor

Pode-se perceber, ao analisar o gráfico apresentado na Figura 52, que o uso e o foco na área da análise de sentimento continua em alta. O uso ou adequações de léxico por um domínio de aplicação teve um aumento nos últimos três anos, até 2012 existiam apenas 4 artigos com esse foco e a partir de 2013 foram encontrados 12 artigos. Atribui-se esse aumento à perceção da necessidade de se levar em consideração elementos do domínio de aplicação para auxiliar na polarização final.

Sobre os artigos com foco nos léxicos, alguns têm como objetivo traduzir léxicos já construídos em inglês para outro idioma, como é possível observar nos trabalhos de Molina-Gonzalez et al. (2013), Martinez-Camara et al. (2014) e Hogenboom et al. (2014). Também existem trabalhos voltados à construção do léxico a partir de textos já polarizados, com o intuito de facilitar a adequação do dicionário com o domínio de aplicação (CRUZ ET AL., 2014; RAO ET AL., 2014).

Dentre os artigos focados em léxico, a prática mais recorrente é a da adequação ou expansão de um léxico já montado para características do domínio em questão. O modelo deste trabalho também permite a adequação do léxico seguindo essa tendência. Pode-se citar alguns trabalhos dessa área, entre eles, Lee, Kim e Yun (2013), Robaldo e Di

Caro (2013), Cho et al. (2014), Wu e Tsai (2014), Dragut et al. (2015), Cotelo et al. (2015) e Park, Lee e Moon (2015).

Após perceber o aumento no foco do desenvolvimento de trabalhos cuja tarefa é construir, manter ou adequar um léxico, optou-se por identificar as técnicas mais utilizadas nos trabalhos a partir de 2013. A Tabela 39 apresenta somente as técnicas que foram utilizadas por pelo menos três trabalhos, distribuídas por ano.

Tabela 39 – Distribuição das técnicas por ano

	2013	**2014**	**2015**	**TOTAL**
SVM	1	10	5	**16**
Redes Neurais	1	5	1	**7**
NLP	1	5	0	**6**
LDA	0	4	1	**5**
Fuzzy	0	4	1	**5**
NB	0	4	0	**4**
Clusterização	0	4	0	**4**
N-Gram	0	4	0	**4**
Coocorrência	2	1	0	**3**

Fonte: Elaborado pelo autor

As técnicas apresentadas na Tabela 39 podem ter sido utilizadas em mais de um trabalho, assim pode-se perceber que o SVM ainda é uma das técnicas mais utilizadas pelas soluções e modelos para análise de sentimento. Pode-se perceber que o uso das redes neurais teve um crescimento importante como técnica para a área. Naive Bayes (NB) pode ser considerada uma das mais utilizadas, mas obteve uma queda na sua utilização em relação à revisão apresentada até meados de 2013.

Ao executar a revisão sistemática inicialmente apresentada no Capítulo 2, pôde-se encontrar o *gap* utilizado como foco para este trabalho no "reaproveitamento de um raciocínio já concretizado, de modo a aproximar-se ao que se tem na aprendizagem humana" aplicado a área de análise de sentimento, de modo que fosse possível explicar a polarização obtida.

A partir desta atualização da revisão sistemática, é possível verificar que a pergunta de pesquisa do trabalho ainda não foi respondida, o que reforça ainda mais a relevância do tema do presente trabalho. Percebe-se com esta nova análise que existem alguns trabalhos que possuem semelhança com a etapa de identificação da polarização de características (em inglês: *features*) ou aspectos (em inglês: *aspects*). Os seguintes trabalhos focam nessa abordagem: Bagheri, Saraee e De Jong

(2014), Dehkharghani (2014), Eltayeby (2014), Kansal e Toshniwal (2014), Lau, Li e Liao (2014), Peñalver-Martinez (2014), Agarwal et al. (2015), Atkinson, Salas e Figueroa (2015), Liu e Chen (2015), Zhao et al. (2015).

A próxima seção tem como objetivo selecionar os artigos que estão mais alinhados com a proposta do modelo da tese para fazer uma análise comparativa analisando as características de cada modelo.

4.5.2 Análise comparativa entre os modelos

A partir do conjunto de artigos apresentados no final da seção anterior, fez-se a leitura completa e foram selecionados os artigos que possuíam características mais próximas aos do modelo de tese.

O trabalho proposto por Agarwal et al. (2015), apresenta uma abordagem interessante para a concepção e atualização de ontologias de domínio para se aplicar à análise de sentimento. Contudo, como o seu foco se diferencia do modelo de tese, o trabalho não faz parte do quadro comparativo. Para a elaboração do quadro comparativo, as seguintes características forma utilizadas como elementos de comparação:

- **armazenamento**: a solução possui uma estratégia para armazenar as classificações já efetuadas?
- **recuperação**: a solução utiliza parte das classificações anteriores como base para novas classificações?
- **explicação**: a solução apresenta uma explicação para a polarização final, ou seja, como se chegou até determinada polarização? e
- **domínio**: a solução é adaptável a domínios distintos, ou seja, trata as ambiguidades do uso dos termos polares?

O Quadro 7 apresenta os artigos coletados e se eles atendem ou não às quatro características apresentadas anteriormente.

Quadro 7– Comparação entre modelos

	Armaz.	**Recuperação**	**Explicação**	**Domínio**
Bagheri, Saraee e De Jong (2014)	Não	Não	Não	Não
Lau, Li e Liao (2014)	Não	Não	Sim	Sim
Peñalver-Martinez (2014)	Não	Não	Sim	Sim
Zhao et al. (2015)	Não	Não	Não	Não

Atkinson, Salas e Figueroa (2015)	Sim	Não	Não	Não
Modelo proposto	**Sim**	**Sim**	**Sim**	**Sim**

Fonte: Elaborado pelo autor

Dos dez trabalhos originalmente resultantes da revisão sistemática para a análise comparativa, cinco foram selecionados para compor o quadro comparativo. Existem dois trabalhos que possuem pelo menos duas características em comum com as características diferenciais da proposta de solução desta pesquisa. Esses dois trabalhos são melhor apresentados nas próximas subseções.

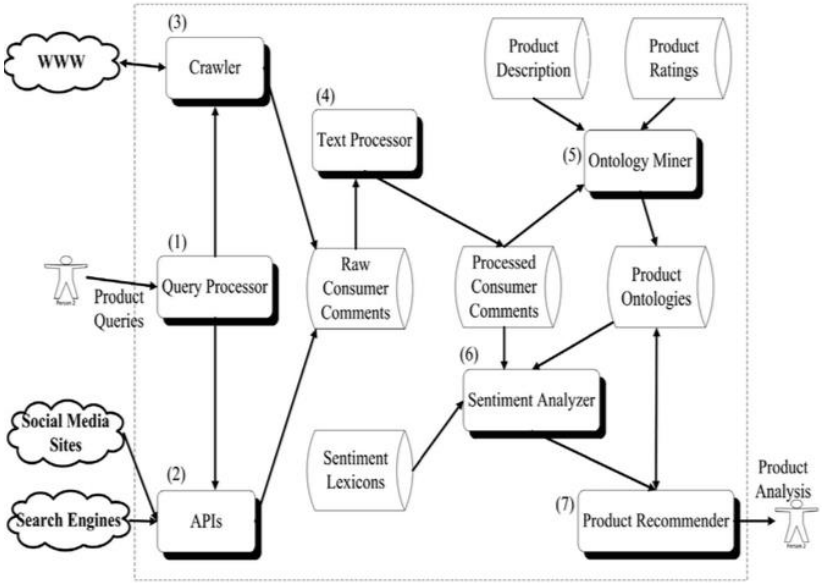
4.5.2.1 Modelo Lau, Li e Liao (2014)

O trabalho proposto por Lau, Li e Liao (2014) apresenta alguns pontos em comum com o modelo proposto. Primeiramente, pode-se comentar que o trabalho utiliza uma ontologia como estrutura formal para modelar o domínio em questão. Além disso, o trabalho também utiliza os conceitos da ontologia como aspectos, ou seja, características de produto ou serviço que se pretende analisar. São vinculados conceitos com os termos polares em questão, também apresentando um grau de relação entre eles.

Como técnica principal para o processo de classificação, esse trabalho utiliza lógica difusa sobre inferências a elementos da ontologia e do texto. Nesse trabalho, apresenta-se também uma etapa de pré-processamento, na qual são utilizadas técnicas de processamento de linguagem natural. É importante mencionar que esse modelo tem como objetivo recomendar produtos e serviços para consumidores finais, dessa forma a análise de sentimento é uma atividade meio e não fim.

A Figura 53, retirada do trabalho original dos autores, demonstra as principais etapas do modelo.

Figura 53 – Modelo proposto por Lau, Li e Liao (2014)



Fonte: Lau Li e Liao (2014), p.83

As Etapas 1, 2 e 3 não são consideradas nesta análise, pois representam processos para extrair os dados das redes sociais ou das ferramentas de *Web 2.0*. Percebe-se que, a partir do conjunto de comentários de produtos e serviços, os textos são submetidos a um processamento utilizando técnicas como: POS *Tagger*, retirada de *stopwords* e *stemmer*.

Após processar os comentários, são buscados nas ontologias de domínios, os conceitos presentes nos textos. Na Etapa 6, utilizam-se os termos polares e verifica-se a sua relação (peso) frente ao texto, submetendo essas informações à etapa de recomendação, etapa esta que está fora do escopo desta tese.

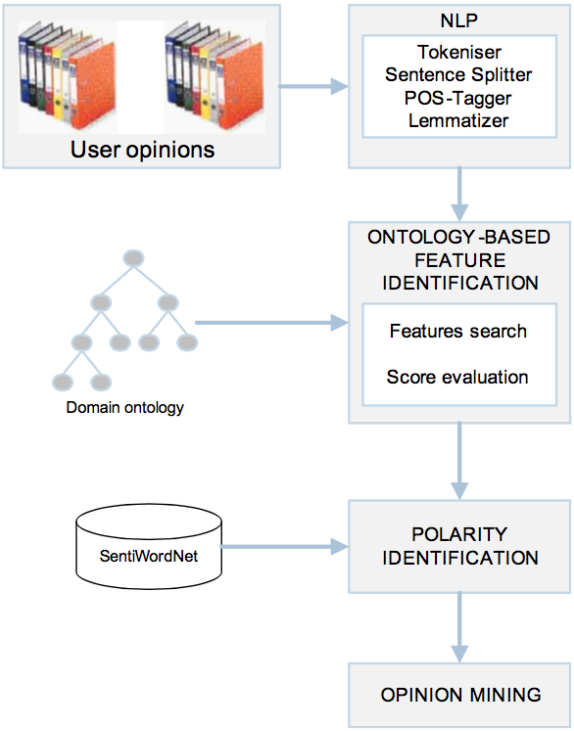
O grande diferencial do modelo proposto neste trabalho em relação ao modelo analisado nesta seção reside no armazenamento de casos já polarizados a fim de permitir o reaproveitamento das classificações passadas, permitindo que essas classificações sejam utilizadas durante processos de adaptação de termos polares por domínios distintos. Outro diferencial é o fato da possibilidade de utilização de casos já polarizados como proposta de solução para novas sentenças ou texto a serem classificados.

A próxima seção apresenta o segundo trabalho que se compara ao modelo proposto.

4.5.2.2 Modelo Peñalver-Martinez et al. (2014)

O trabalho de Peñalver-Martinez et al. (2014) apresenta uma proposta bastante similar ao trabalho de Lau, Li e Liao (2014), mas com foco na própria análise de sentimento. A Figura 54 foi extraída do trabalho original dos autores.

Figura 54 – Modelo proposto por Peñalver-Martinez et al. (2014)



Fonte: Peñalver-Martinez et al. (2014), p. 5999

Os *reviews* dos usuários são submetidos ao modelo e na sua primeira etapa tem-se o pré-processamento do texto utilizando ferramental da área do processamento de linguagem natural. O próximo passo envolve a busca na ontologia de domínio pelos conceitos (características dos serviços ou produtos) existentes no texto.

De posse dos conceitos envolvidos, busca-se no *SentiWordNet* os termos polares presentes no texto. A partir da contabilização dos termos polares por características, é calculada a polarização final e apresentada para o usuário.

O modelo desta tese se diferencia desse modelo exatamente da mesma forma que apresentada na Seção 4.5.2.1, ou seja, esse modelo não utiliza casos passados de forma a adaptar termos polares ambíguos e nem reutiliza classificações passadas como proposta de solução.

Na visão dos autores, uma das principais contribuições do trabalho é a ontologia gerada a partir da *MovieOntology*, que pode ser utilizada para novos trabalhos desse domínio na área da análise de sentimento.

A seção seguinte apresenta as considerações finais do presente capítulo.

4.6 CONSIDERAÇÕES FINAIS

O presente capítulo teve como objetivo avaliar o modelo inicial da tese, proposto após o levantamento do estado da arte e do referencial teórico. Para isso, foi concebido um recorte do modelo para atestar a viabilidade de trabalhar com a construção de uma árvore de conceitos e instâncias vindas de uma ontologia combinadas com termos polares vindos de um léxico de sentimento. Também se objetivou para esse primeiro experimento a possibilidade de armazenar os casos já processados a fim de utilizá-los em novas classificações (para isso utilizou-se RBC).

Após a implementação do protótipo baseado no recorte da proposta de solução, avaliou-se o seu uso aplicado ao domínio de comentários sobre câmeras digitais. O mesmo processo foi reproduzido utilizando SVM e NB, em que o modelo proposto obteve resultado superior após o processo de adequação da ontologia de domínio e do léxico utilizado.

Ao finalizar essa avaliação inicial do recorte do modelo proposto, percebeu-se que existiam pontos de melhoria. Tal constatação foi importante, pois permitiu a definição dos passos seguintes para o modelo em questão.

Percebeu-se que existia um problema no tratamento de termos ambíguos que podem ter uma conotação (polarização) para um domínio e outra completamente diferente para outro domínio. Desta forma, optou-se pela definição de um método para evoluir a base de conhecimento, trazendo todas as particularidades de mudança nas polarizações originais de termos do léxico geral, proposto por Hu e Liu (2004), para dentro da ontologia, tornando essa modificação uma característica do domínio.

Sabendo que o processo de adequação do léxico é algo bastante trabalhoso, construiu-se um processo automatizado para facilitar a

identificação e a adequação de termos polares em um novo domínio utilizando uma base de treinamento.

O mesmo processo de avaliação foi efetuado levando em consideração um novo domínio, dessa vez, de comentários de filmes vendidos no *site* da Amazon. A partir do ferramental construído para adequar as bases de conhecimento, chegou-se a definição da última etapa do modelo proposto, a adequação.

Foram construídas três formas distintas para adaptar um termo ambíguo baseado no seu contexto e nos casos já armazenados. A que obteve melhor resultado foi a técnica baseada no consumo de dados e de informações presentes nas matrizes SVD. Entretanto, o processo de decomposição matricial (SVD) é custoso e deve sofrer atualizações de tempos em tempos. Caso esses requisitos não sejam viáveis, o método baseado na frequência dos termos pode ser utilizado com resultados satisfatórios.

Considerando o modelo proposto final, foram efetuadas mais avaliações baseadas nos dois domínios já trabalhados, comparando os resultados com o SVM e NB e obtendo resultados superiores.

Por fim, optou-se por atualizar a revisão sistemática e apresentar novamente o estado da arte. Verificou-se com isso que a lacuna que esse trabalho procura atender continua aberta. Essa conclusão pode ser percebida por meio da análise de trabalhos com abordagens similares como os apresentados nas Seções 4.5.2.1 e 4.5.2.2.

Observou-se que, principalmente as etapas de recuperação e adaptação, do modelo proposto, não estão sendo ainda tratadas em outros trabalhos neste contexto. Para isso, atribui-se o uso do RBC e das matrizes SVD utilizando como base os casos armazenados.

O RBC (base de casos) contribui de duas maneiras para o modelo da tese: (1) ao recuperar um caso inteiro como proposta de solução, bem como, (2) na etapa de adaptação, na qual é possível, a partir dos termos ambíguos, utilizar parte de classificações passadas para auxiliar na polarização de um novo caso.

O próximo capítulo tem como objetivo apresentar as conclusões e os trabalhos futuros.

5 CONCLUSÕES E TRABALHOS FUTUROS

Este capítulo tem como objetivo apresentar as conclusões obtidas durante o processo de desenvolvimento deste trabalho por meio das etapas de avaliação e análise dos resultados do modelo proposto.

Além das conclusões, também são apresentados os trabalhos futuros. A produção bibliográfica gerada durante o processo de doutoramento pode ser consultada no Apêndice deste trabalho.

5.1 CONCLUSÕES

Este trabalho está focado na área de classificação de textos, mais precisamente na área de análise de sentimento. Por meio de revisões sistemáticas da literatura, pôde-se perceber que a área está sendo bastante pesquisada, ou seja, existem vários trabalhos sendo produzidos. A partir da problemática, pôde-se observar uma série de oportunidades de pesquisa, assim foi possível identificar os pontos principais que o modelo da tese deveria tratar.

Para obter uma análise mais adequada dos sentimentos presentes em texto, deve-se levar em consideração elementos do domínio de aplicação. Além disso, deve-se aprender com classificações passadas, permitindo uma melhoria em novas classificações. Outro ponto que foi explicitado na problemática é que, em muitos casos, conhecer o motivo de determinada classificação é tão importante quanto o próprio resultado em si para o apoio a decisão.

Também foi identificado que a negação deve ser tratada pelas soluções de análise de sentimento e, principalmente, para as soluções que são baseadas em léxico. Além disso, deve-se existir um tratamento adequado dos termos ambíguos, pois eles podem trazer muito problemas para a classificação do texto.

Até a elaboração deste trabalho, nenhum outro havia combinado todos esses elementos em uma única solução. Essa informação foi obtida por meio de duas revisões sistemáticas realizadas, sendo uma no início desta pesquisa e outra ao final. O modelo proposto aqui buscou combinar vários recursos e técnicas para chegar a um aumento na acurácia dos textos classificados, além de permitir que seja apresentada uma explicação do motivo de atingir determinada classificação.

O uso do RBC permitiu que os textos (casos) já classificados fossem armazenados e reutilizados como proposta de solução para novos textos a serem classificados. Outra contribuição que o RBC deu a este trabalho foi a base de casos polarizados, que é gerada à medida que os textos são classificados. Nessa base de casos, são armazenados os

casos originais, a sua polarização final e a explicação de como se chegou até a polarização. Essas informações podem ser utilizadas como base para sistemas baseados em conhecimento, como, por exemplo, para apresentar como um produto ou serviço é visto pelo conjunto de termos já classificados. Também é possível verificar como a explicação é modificada ao longo do tempo, levando em consideração a data em que o texto foi publicado ou inserido na base.

A partir da base de casos já processados, é possível gerar as matrizes SVD que se apresentam como uma forma de identificar o contexto em que um termo polar é utilizado, possibilitando verificar se o termo, para aquele texto, tem conotação positiva ou negativa. Esta é a base para a etapa de adaptação do modelo proposto.

Ainda sobre o contexto do texto a ser classificado, o uso de ontologias de domínio permitiu apresentar as características dos serviços e dos produtos a partir das classes e instâncias existentes no texto a ser classificado, preservando a lógica de ligação das classes. É por meio dessas ligações (relacionamentos) que são elaboradas, no contexto da tese, as chamadas árvores de sentimento, que podem ser entendidas como uma estrutura de grafo que apresenta os conceitos e instâncias das ontologias ligados pelas suas relações e combinados com os termos polares. Sendo assim, a árvore de sentimento pode ser utilizada como explicação para a classificação.

Para obter os termos polares, foi utilizado um léxico de sentimento que, no contexto da tese, foi adaptado para cada domínio a partir de um processo que verifica os textos já polarizados na base de casos.

Pelo resultado dos testes efetuados, pôde-se perceber que a base do modelo, o seu núcleo, é a construção das árvores de sentimento, se baseando em conceitos de uma ontologia de domínio e na contabilização de termos polares a partir de um léxico. Ao adicionar qualquer uma das novas etapas de maneira isolada, obteve-se uma maior acurácia. Pôde-se perceber que o uso de RBC e o tratamento de negação apresentaram uma melhoria inicial para a acurácia total do modelo, mas o uso das etapas de adequação do léxico e adaptação apresentaram uma melhoria muito maior. Agora, quando se combinam todas essas etapas para a solução final, obtém-se a maior acurácia. O que demonstra que combinadas, todas essas etapas apresentam uma contribuição para o modelo final da tese.

A orquestração de técnicas e ferramentas possibilitaram a obtenção de resultados superiores em termos de acurácia quando comparados com algoritmos tradicionais, tais como, SVM e NB, além

de permitir a apresentação de uma explicação para a classificação em questão.

Tendo em vista os pontos aqui levantados, afirma-se que a pergunta de pesquisa foi respondida, ou seja, é possível recuperar e armazenar um conhecimento representado na forma de uma árvore de sentimento a fim de auxiliar na tarefa intensiva de classificação com foco na análise de sentimento.

A próxima seção tem como objetivo apresentar os possíveis trabalhos futuros no tema.

5.2 TRABALHOS FUTUROS

O desenvolvimento deste trabalho possibilitou a identificação de vários pontos de evolução para a área de análise de sentimentos bem como para o próprio modelo da tese.

Esta proposta utiliza ontologias para permitir que as análises efetuadas utilizem elementos do domínio de aplicação. Considerando as ontologias de domínio, seria importante desenvolver pesquisas para a construção e evolução das ontologias a partir de novos *reviews* que venham a ser submetidos ao processo de classificação.

Ainda sobre as ontologias, seria adequado desenvolver um método para identificação automática do contexto do texto a ser classificado, permitindo que uma ontologia mais apropriada pudesse ser instanciada e utilizada durante o processo de classificação sem a necessidade de informar o domínio antes do processo.

Um dos grandes benefícios do uso do modelo da tese é o reaproveitamento de casos (textos) já polarizados para decidir se um caso pode ser apresentado como proposta de solução para um novo texto. Para isso utilizou-se um *threshold*. Seria importante desenvolver estudos sobre novas maneiras para calcular esse valor, de modo a aproveitar ainda mais os benefícios do RBC no modelo de tese.

Um ponto que pode trazer benefícios para o uso do modelo proposto é a construção de um módulo anterior à etapa de pré-processamento que identifique se o texto em questão é subjetivo ou objetivo, eliminando dessa forma textos que não são passíveis de classificação.

Outro aspecto do modelo passível de evolução é o tratamento de sentenças com negação. A proposta atual utiliza um método bastante simples, mas pode-se pensar na utilização de regras mais elaboradas, combinadas com elementos do processo de adaptação.

Os testes efetuados para a avaliação do modelo da tese foram baseados na língua inglesa, incluindo textos (*reviews*), ontologias e o

léxico. O modelo está preparado para suportar classificações em qualquer idioma, desde que a base de conhecimento seja construída para isso. Sendo assim, podem-se desenvolver mecanismos para traduzir a base de conhecimento para outros idiomas. Ainda nessa linha, seria interessante desenvolver um método para identificar o idioma do texto antes da sua classificação, dessa forma facilitaria a instanciação dos recursos para a nova classificação.

Finalmente, o cálculo para gerar o valor final da polarização utiliza apenas -1 para termos negativos e +1 para termos positivos. Existem trabalhos que buscam pesos diferentes para termos em determinados domínios, sendo esta uma pesquisa que pode promover benefícios para o modelo da tese e para o valor da polarização final.

REFERÊNCIAS

AAMODT, Agnar; PLAZA, Enric. **Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches**. 1994. AI Communications. IOS Pres. Vol 7.

ABBASI, A.; CHEN, H.; SALEM, A. Sentiment analysis in multiple languages- Feature selection for opinion classification in Web forums. **ACM Transactionson Information Systems**, v. 26, n. 3, p. 12-46, 2008.

ADOLPHO, Conrado. **Os 8Ps do Marketing Digital: O guia estratégico de Marketing Digital**. São Paulo: NovatecEditora, 2011.

AGARWAL, Basant et al. Sentiment Analysis Using Common-Sense and Context Information. **Computational intelligence and neuroscience**, v. 2015, 2015.

AISOPOS, Fotis. et al. **Content vs. Context for Sentiment Analysis: a comparative analysis over microblogs**. Proceedings of the 23rd ACM conference on Hypertext and social media. P. 187-196, New York, 2012.

AMBINDER, D. M; MARCONDES, C. H. **As Potencialidades da Web Semântica e Web 2.0 para a Ciência da Informação e os novos Formatos de Publicações Eletrônicas para a Pesquisa Acadêmico-científica**. Revista EDICIC, v.1, n.4, p.342-362, Out./Dez. 2011.

ANTONINI, A. et al. **Tracking and Analyzing TV Content on the Web through Social and Ontological Knowledge**. In: EuroITV'13 ACM, June 24–26, 2013, Como, Italy. 2013.

AREL, Itamar; ROSE, Derek C.; KARNOWSKI, Thomas P. Deep machine learning-a new frontier in artificial intelligence research [research frontier]. **Computational Intelligence Magazine, IEEE**, v. 5, n. 4, p. 13-18, 2010.

ATKINSON, John; SALAS, Gonzalo; FIGUEROA, Alejandro. Improving opinion retrieval in social media by combining features-based coreferencing and memory-based learning. **Information Sciences**, v. 299, p. 20-31, 2015.

BAGHERI, Ayoub; SARAEE, Mohamad; DE JONG, Franciska. ADM-LDA: An aspect detection model based on topic modelling using the structure of review sentences. **Journal of Information Science**, v. 40, n. 5, p. 621-636, 2014.

BALAHUR, Alexandra et al. **EmotiNet: a knowledge base for emotion detection in text built on the appraisal theories**. In: Natural Language Processing and Information Systems. Springer Berlin Heidelberg, 2011. p. 27-39.

BALANCIERI, Renato. **Um Método Baseado em Ontologias para Explicitação de Conhecimento derivado da Análise de Redes Socias de um domínio de aplicação**. (Tese) Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. Universidade Federal de Santa Catarina. Florianópolis, 2010.

BALDONI, Matteo et al. **From tags to emotions: Ontology-driven sentiment analysis in the social semantic web**. Intelligenza Artificiale, v. 6, n. 1, p. 41-54, 2012.

BANEA, C.; MIHALCEA, R.; WIEBE, Janyce. **A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources**. International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, 2008.

BEIL, Florian; ESTER, Martin; XU, Xiaowei. Frequent term-based text clustering. In: **Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining**. ACM, 2002. p. 436-442.

BELLINI, F. et al. **The Role of the Prosumer in Social Networks and the Sentiment Analysis for the Customer Experience Management**. cersi.it, 2012.

BENGIO, Yoshua. Deep Learning of Representations for Unsupervised and Transfer Learning. In: **ICML Unsupervised and Transfer Learning**. 2012. p. 17-36.

BEPPLER, F. D. **Emprego de RBC para Recuperação Inteligente de Informação**. Dissertação. Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarina. 2002.

BEPPLER, Fabiano D. **Um Modelo para Recuperação e Busca de Informação Baseado em Ontologia e no Círculo Hermenêutico**. (Tese) Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. Universidade Federal de Santa Catarina. Florianópolis, 2008.

BERGMANN, Ralph; SCHAAF, Martin. **Structural Case-Based Reasoning and ontology-based knowledge management: A perfect match?**. J. UCS, v. 9, n. 7, p. 608-626, 2003.

BIANCHINI, A. R. Arquitetura de redes neurais para o reconhecimento facial baseado no neocognitron [dissertação]. **São Carlos: Universidade Federal de São Carlos**, 2001.

BOIY, Erik; MOENS, Marie-Francine. A machine learning approach to sentiment analysis in multilingual Web texts. **Information retrieval**, v. 12, n. 5, p. 526-558, 2009.

BONA, Cristina. **Avaliação de processos de software: um estudo de caso em XP e Iconix**. 2002. 122 f. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis, 2002.

BORST, Willem Nico. **Construction of engineering ontologies for knowledge sharing and reuse**. Tese de Doutorado, UniversiteitTwente, 1997.

BORTH, D. et al. **SentiBank: Large-Scale Ontology and Classifiers for Detecting Sentiment and Emotions in Visual Content**. MM'13 ACM, October 21–25, 2013, Barcelona, Spain, 2013.

BOYD, D. M.; ELLISON, N. B. **Social network sites: Definition, history, and scholarship**. Journal of Computer-Mediated Communication, 13, p .210-230.2008.

BRACHMAN, Ronald J.; LEVESQUE, Hector J. **Knowledge representation and reasoning**. Morgan Kaufmann Publishers, 2004.

BUCHE, A.; CHANDAK, M.B.; ZADGAONKAR, A. **Opinion Mining and Analysis: A Survey**. International Journal on Natural Language Computing (IJNLC) v. 2, n .3, 2013.

CAMBRIA, E. et al. **Sentic Computing for social media marketing**. Multimed Tools Appl. v.59, p.557–577, 2012A.

CAMBRIA, E. et al. **Sentic PROMs: Application of sentic computing to the development of a novel unified framework for measuring health-care quality**. Expert Systems with Applications. v.39, p.10533–10543, 2012B.

CAO, Yanfang; ZHANG, Pu; XIONG, Anping. **Sentiment Analysis Based on Expanded Aspect and Polarity-Ambiguous Word Lexicon**. International Journal of Advanced Computer Science and Applications (IJACSA), v. 6, n. 2. 2015.

CARPINETO, Claudio et al. **A survey of web clustering engines**. ACM Computing Surveys (CSUR), v. 41, n. 3, p. 17, 2009.

CECI, Flavio et al. **Towards a semi-automatic approach for ontology maintenance**. In: CONTECSI INTERNATIONAL CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGY MANAGEMENT, 7., 2010, São Paulo. Anais… São Paulo: USP, 2010.

CECI, Flavio; PIETROBON, Ricardo; GONÇALVES, Alexandre Leopoldo. **Turning Text into Research Networks: Information Retrieval and Computational Ontologies in the Creation of Scientific Databases**. PlosOne, v. 7, p. e27499, 2012.

CECI, Flavio; WOSZEZENKI, Cristiane Raquel; GONÇALVES, Alexandre Leopoldo. O uso de anotações semânticas e ontologias para a classificação de documentos. **International Journal of Knowledge Engineering and Management** (IJKEM), v. 3, n. 5, p. 1-14, 2014.

CHAVES, MarcirioSilveira. **Uma MetodologiaparaConstrução de Geo-Ontologias**. 2009. (Tese de Doutoramento) – Programa de

DoutoramentoemInformática da Universidade de Lisboa, Universidade de Lisboa, Portugal, 2009.

CHANDRAN, S.; MURUGAPPAN, S. **A Review on Opinion Mining from Social Media Networks**. European Journal of Scientific Research, v. 89, n. 3, p. 430-440, 2012.

CHANDRASEKARAN, B.; JOSEPHSON, J. R. What are Ontologies, and Why Do We Need Them?. IEEE Intelligent Systems, IEEE, p. 20-26, 1999.

CHEN, Jingnian et al. Feature selection for text classification with Naïve Bayes. **Expert Systems with Applications**, v. 36, n. 3, p. 5432-5435, 2009.

CHEN, C.; CHEN, Z.; WU, C. **An Unsupervised Approach for Person Name Bipolarization Using Principal Component Analysis**. IEEE Transactions of Knowledge and Data Engineering, v. 24, n. 11, 2012.

CHEN, H.; et al. Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. **Frontiers in Physiology**, v. 4, n. January, p. 1–6, 2013.

CHEONG, Marc; LEE, Vincent CS. **A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter**. Information Systems Frontiers, v. 13, n. 1, p. 45-59, 2011.

CHO, Heeryon et al. Data-driven integration of multiple sentiment dictionaries for lexicon-based sentiment classification of product reviews. **Knowledge-Based Systems**, v. 71, p. 61-71, 2014.

CIMIANO, Philipp; VOLKER, Johanna. **Text2Onto: a framework for ontology learning and data-driven change discovery**. In: INTERNATIONAL CONFERENCE ON APPLICATIONS OF NATURAL LANGUAGE TO INFORMATION SYSTEMS (NLDB), 10., 2005, Alicante, Spain. Proceedings... Alicante, Spain: Springer, 2005.

CORCHO, O.; GOMEZ-PEREZ, A. **A roadmap to ontology specification languages**. Knowledge Engineering and Knowledge Management. Methods, Models, and Tools, Springer Berlin / Heidelberg, p. 80–96, 2000.

COTELO, J. M. et al. A modular approach for lexical normalization applied to Spanish tweets. **Expert Systems with Applications**, v. 42, n. 10, p. 4743-4754, 2015.

CRUZ, F. L. et al. 'Long autonomy or long delay?'The importance of domain in opinion mining. **Expert Systems with Applications**, v. 40, n. 8, p. 3174-3184, 2013.

CRUZ, Fermín L. et al. Building layered, multilingual sentiment lexicons at synset and lemma levels. **Expert Systems with Applications**, v. 41, n. 13, p. 5984-5994, 2014.

DAVE, Kushal et al. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: **Proceedings of the 12th international conference on World Wide Web**. ACM, 2003. p. 519-528.

DEHKHARGHANI, Rahim et al. Sentimental causal rule discovery from Twitter. **Expert Systems with Applications**, v. 41, n. 10, p. 4950-4958, 2014.

DENG, H. et al. **Exploring and Inferring User-User Pseudo-Friendship for Sentiment Analysis with Heterogeneous Networks**. SIAM Conference on Data Mining (SDM13), Texas, 2013.

DEVROYE, Luc; WAGNER, T. J. Distribution-free inequalities for the deleted and holdout error estimates. **Information Theory, IEEE Transactions on**, v. 25, n. 2, p. 202-207, 1979.

DONG, Ruihai et al. Opinionated Product Recommendation. In: **Case-Based Reasoning Research and Development**. Springer Berlin Heidelberg, p. 44-58. 2013a.

DONG, Ruihai et al. **Mining Features and Sentiment from Review Experiences**. In: Case-Based Reasoning Research and Development. Springer Berlin Heidelberg, p. 59-73. 2013b.

DRAGUT, Eduard C. et al. Polarity Consistency Checking for Domain Independent Sentiment Dictionaries. **Knowledge and Data Engineering**, IEEE Transactions on, v. 27, n. 3, p. 838-851, 2015.

DRUZIANI, Cássio Frederico Moreira; KERN, Vinicius Medina; CATAPAN, Araci Hack. A Gestão e a Engenharia do Conhecimento Aliadas na Modelagem do Conhecimento: Análise Sistêmica CESM e Contextual CommonKADS de um Repositóriona Web. **Perspectivas em Gestão & Conhecimento**, v. 2, n. 1, p. 194-217, 2012.

DURIC, Adnan; SONG, Fei. Feature selection for sentiment analysis based on content and syntax models. In: **Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis**. Association for Computational Linguistics, 2012. p. 96-103.

EFRON, Miles. **Using cocitation information to estimate political orientation in web documents**. Knowledge and Information Systems, v. 9, n. 4, p. 492-511, 2006.

EL SAYED, Ahmad; HACID, Hakim. **A hybrid approach for taxonomy learning from text**. COMPSTAT 2008. p. 255-266, 2008.

ELTAYEBY, Omar; MOLNAR, Peter; GEORGE, Roy. Measuring the Influence of Mass Media on Opinion Segregation through Twitter. **Procedia Computer Science**, v. 36, p. 152-159, 2014.

ESULI, Andrea. **Automatic Generation of Lexical Resources for Opinion Mining: Models, Algorithms and Applications**. Tese de doutorado em Engenharia da Informação. Universitàdi Pisa. Itália, 2008.

FAN, Teng-Kai; CHANG, Chia-Hui. Sentiment-oriented contextual advertising. **Knowledge and Information Systems**, v. 23, n. 3, p. 321-344, 2010.

FELDMAN, Ronen. Techniques and applications for sentiment analysis. **Communications of the ACM**, v. 56, n. 4, p. 82-89, 2013.

FENG, Shi et al. Extracting common emotions from blogs based on fine-grained sentiment clustering. **Knowledge and information systems**, v. 27, n. 2, p. 281-302, 2011.

FENSEL, Dieter. **Ontologies: silver bullet for knowledge management and electronic commerce**. Berlin: Springer-Verlag, 2001.

FERNÁNDEZ, Ronald T.; LOSADA, David E. Effective sentence retrieval based on query-independent evidence. **Information Processing & Management**, v. 48, n. 6, p. 1203-1229, 2012.

FERREIRA, Aurélio Buarque de Holanda. **Míni Aurélio**. O Dicionário da Língua Portuguesa. Curitiba: Editora Positivo, 2009.

FIALHO, Francisco, A. P. **Psicologia das Atividades Mentais**: Introdução às Ciências da Cognição. Florianópolis: Editora Insular, 2011.

FORTUNA, Blaž; LAVRAČ, Nada; VELARDI, Paola. **Advancing topic ontology learning through term extraction**. Springer Berlin Heidelberg, 2008.

FRANZ, Marie-Louise von. **Tipologia de Jung**. São Paulo: Cultrix, 2003.

FREITAS, Alex A. A genetic programming framework for two data mining tasks: classification and generalized rule induction. In: **Genetic Programming 1997: Proc 2nd Annual Conf.** Morgan Kaufmann, 1997. p. 96-101.

FREITAS, Frederico Luiz G. de. **Ontologias e a Web Semântica**. In: Renata Vieira; Fernando Osório. (Org.). **Anais do XXIII Congresso da Sociedade Brasileira de Computação**. Volume 8: Jornada de Mini-Cursos em Inteligência Artificial. Campinas: SBC, 2003, v. 8, p. 1-52.

FREITAS, Larissa A.; VIEIRA, Renata. Ontology based feature level opinion mining for portuguese reviews. In: Proceedings of the **22nd international conference on World Wide Web companion**. International World Wide Web Conferences Steering Committee, 2013a. p. 367-370.

FREITAS, L. A.; VIEIRA, R. Comparing Portuguese Opinion Lexicons in Feature-Based Sentiment Analysis. **IJCLA.** v. 4, n. 1, p. 147-158. JAN-JUN, 2013b.

GACITUA, Ricardo; SAWYER, Pete; RAYSON, Paul. A flexible framework to experiment with ontology learning techniques. In: RESEARCH AND DEVELOPMENT IN INTELLIGENT SYSTEMS, 24., 2007, London . **Proceedings...** London: Springer, 2007. p. 153-166.

GAILLARD, Emmanuelle et al. Case-Based Reasoning on E-Community Knowledge. In: **Case-Based Reasoning Research and Development**. Springer Berlin Heidelberg, 2013. p. 104-118.

GARCIA-CONSTANTINO, M. F. **On The Use Of Text Classification Methods For Text Summarisation**. Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy, 2013

GO, Alex; BHAYANI, Richa; HUANG, Lei. **Twitter Sentiment Classification using Distant Supervision**. Technical report, Stanford Digital Library Technologies Project. 2009.

GODBOLE, N.; SRINIVASAIAH, M.; SKIENA, S. **Large-Scale Sentiment Analysis for News and Blogs**. International Conference on Weblogs and Social Media (ICWSM). Colorado. 2007.

GÓMEZ-PÉREZ, Asunción; FERNÁNDEZ-LÓPEZ, Mariano; CORCHO, Oscar. **Ontology Engineering – with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web**. London: Springer-Verlag, 2004.

GONÇALVES, Alexandre Leopoldo. **Um modelo de descoberta de conhecimento baseado na correlação de elementos textuais e expansão vetorial aplicado à Engenharia e Gestão do Conhecimento**. Tese apresentada ao Programa de Pós-Graduação em Engenharia de Produção da Universidade Federal de Santa Catarinacomo requisito parcial para obtenção do grau de Doutor em Engenharia de Produção. Florianopolis, 2006.

GRANITZER, Michael et al. **Automated ontology learning and validation using hypothesis testing**. Advances in Soft Computing, Berlin, v. 43, p. 130-135, 2007.

GLOROT, Xavier; BORDES, Antoine; BENGIO, Yoshua. Domain adaptation for large-scale sentiment classification: A deep learning approach. In: **Proceedings of the 28th International Conference on Machine Learning (ICML-11)**. 2011. p. 513-520.

GRUBER, T. **A translation approach to portable ontology specification**. Knowledge Acquisition, v. 5, n. 2, pag. 199-220, 1993.

GRUBER, Thomas R. Toward principles for the design of ontologies used for knowledge sharing?. **International journal of human-computer studies**, v. 43, n. 5, p. 907-928, 1995.

GUARINO, Nicola. **Formal Ontology and Information Systems**. FOIS´98. IOS Press, Amisterdan, 1998.

GUERRA, P. H. C. et al. **From Bias to Opinion: a Transfer-Learning Approach to Real-Time Sentiment Analysis**. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. p. 150-158, 2011.

GUNAWARDENA, S; WEVER, R. O. **Applying CBR Principles to Reason without Negative Exemplars**. Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference. p. 405-408, 2012.

HATZIVASSILOGLOU, Vasileios; MCKEOWN, Kathleen R. **Predicting the semantic orientation of adjectives**. Proceedings of the 35th annual meeting on Association for Computational Linguistics, 1997.

HAYKIN, S. S. Redes neurais artificiais: princípio e prática. **2ª Edição, Bookman**, São Paulo, Brasil, 2001.

HE, Yulan; ALANI, Harith; ZHOU, Deyu. **Exploring English Lexicon Knowledge for Chinese Sentiment Analysis**. In: CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing, China, 2010.

HEITMANN, Benjamin; HAYES, **Conor. Enabling Case-Based Reasoning on the Web of Data**. In: Workshop Proceedings. 2010. p. 131.

HOGENBOOM, Alexander et al. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. **Decision support systems**, v. 62, p. 43-53, 2014.

HU, X. et al. **Exploiting Social Relations for Sentiment Analysis in Microblogging**. the 6th ACM International Conference on Web Search and Data Mining (WSDM 2013), Roma, 2013.

HU, Minqing; LIU, Bing. Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004. p. 168-177.

HU, Yuheng; WANG, Fei; KAMBHAMPATI, Subbarao. Listening to the crowd: automated analysis of events via aggregated twitter sentiment. In: **Proceedings of the Twenty-Third international joint conference on Artificial Intelligence**. AAAI Press, 2013. p. 2640-2646.

HUANG, Z.; QIU, Yuhui. **A multiple-perspective approach to constructing and aggregating Citation Semantic Link Network**. Future Generation Computer Systems n.26, p.400–407, 2010.

JEBASEELI, A.; KIRUBAKARAN, E. **Opinion Mining of M Learning Reviews using Soft Computing Techniques**. International Journal of Computer Applications, v. 54, n. 15, p. 44-48, 2012.

JI, Chuang et al. **A Case-Based Reasoning System for Residual Value Risk in Public-Private Partnership Projects**. In: ICCREM 2013@ Construction and Operation in the Context of Sustainability. ASCE. p. 680-692. 2013.

JUNIOR, D. T. et al. **Sistema de Raciocínio Baseado em Casos para Recomendação de Programa Alimentar**. RESI – Revista Eletrônica de Sistemas de Informação, n.3, 2006.

KAISER, Carolin; SCHLICK, Sabine; BODENDORF, Freimut. Warning system for online market research–identifying critical

situations in online opinion formation. **Knowledge-Based Systems**, v. 24, n. 6, p. 824-836, 2011.

KAMPS, Jaap; et al. Using wordnet to measure semantic orientation of adjectives. In: Proceedings of the **4th International Conference on Language Resources and Evaluation** (LREC 2004). 2004. p. 1115-1118.

KANAYAMA, H.; NASUKAWA, T. **Fully automatic lexicon expansion for domain-oriented sentiment analysis**. Proceedings of the Conference on Empirical Methods in Natural Language Processing – EMNLP 2006, p. 355, 2006.

KANG, Hanhoon; YOO, Seong Joon; HAN, Dongil. Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews. **Expert Systems with Applications**, v. 39, n. 5, p. 6000-6010, 2012.

KANSAL, Hitesh; TOSHNIWAL, Durga. Aspect based Summarization of Context Dependent Opinion Words. **Procedia Computer Science**, v. 35, p. 166-175, 2014.

KASTER, D. S.; MEDEIROS, C. B.; ROCHA, H. V. **Aplicação de Raciocínio Baseado em Casos a Sistemas de Apoio à Decisão Ambienta**l. IC-UNICAMP, Campinas, São Paulo, Brasil. 2000.

KAZAMA, J; TSUJII, J. **Maximum entropy models with inequality constraints**: A case study on text categorization. Machine Learning, v. 60, n. 1-3, p. 159-194, 2005.

KIM, Soo-Min; HOVY, Eduard. **Determining the Sentiment of Opinions**. The 20th International Conference on Computational Linguistics. In Proceedings of COLING. Suiça, 2004.

KOLODNER, J. L. **Case-Based Reasoning**. Morgan Kaufmann Pub., Inc. Californian, United States, 1993.

KONCHADY, Manu. **Text mining application programming**. Massachusetts: Charles River Media, 2006.

KONTOPOULOS, E. et al. **Ontology-based sentiment analysis of twitter posts. Expert Systems with Applications.** v. 40, 4065–4074, 2013.

KOTSIANTIS, Sotiris B.; ZAHARAKIS, I. D.; PINTELAS, P. E. **Supervised machine learning**: A review of classification techniques. 2007.

KU, Lun-Wei; LIANG, Yu-Ting; CHEN, Hsin-Hsi. **Tagging heterogeneous evaluation corpora for opinionated tasks**. Proceedings of the Fifth International Conference on Language Resources and Evaluation, pages 667-670, Genoa, Italy, May 24-26, 2006.

LACHMAN, Gary. **Jung the Mystic: The Esoteric Dimensions of Carl Jung's Life and Teachings**. Penguin, 2010.

LANE, Peter CR; CLARKE, Daoud; HENDER, Paul. On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data. **Decision Support Systems**, v. 53, n. 4, p. 712-718, 2012.

LARMAN, Craig. **Utilizando UML e padrões: uma introdução à análise e ao projeto orientado a objeto**. Bookman, Porto Alegre, 2000.

LAU, Raymond YK; LI, Chunping; LIAO, Stephen SY. Social analytics: Learning fuzzy product ontologies for aspect-oriented sentiment analysis. **Decision Support Systems**, v. 65, p. 80-94, 2014.

LEE, Kong-Joo; KIM, Jee-Eun; YUN, Bo-Hyun. Extracting Multiword Sentiment Expressions by Using a Domain-Specific Corpus and a Seed Lexicon. **ETRI Journal**, v. 35, n. 5, p. 838-848, 2013.

LEVY, Moria. **WEB 2.0 implications on knowledge management**. Journal of Knowledge Management, Vol 13, p. 120-134, 2009.

LI, Sheng-Tun; TSAI, Fu-Ching. A fuzzy conceptualization model for text mining with application in opinion polarity classification. **Knowledge-Based Systems**. v.39 p.23–33, 2013.

LI, Lin; XIA, Yunqing; ZHANG, Pengzhou. An Unsupervised Approach to Sentiment Word Extraction in Complex Sentiment

Analysis. **International Journal of Knowledge and Language Processing**. Vol. 2, p.40-52, 2011.

LI, Weiyuan; XU, Hua. Text-based emotion classification using emotion cause extraction. **Expert Systems with Applications**, v. 41, n. 4, p. 1742-1749, 2014.

LIU, Bing. **Sentiment analysis and subjectivity**.A chapter in Handbook of Natural Language Processing, Second edition. 2010a.

LIU, Bing. **SentimentAnalysis: A Multi-Faceted Problem**. Invited paper, IEEE Intelligent Systems, Vol. 25, p. 76-80, 2010b.

LIU, Li-zhen et al. Generating domain-specific affective ontology from Chinese reviews for sentiment analysis. **Journal of Shanghai Jiaotong University (Science)**, v. 20, p. 32-37, 2015.

LIU, Shuhua Monica; CHEN, Jiun-Hung. A multi-label classification based approach for sentiment classification. **Expert Systems with Applications**, v. 42, n. 3, p. 1083-1093, 2015.

LIU, Ying et al. Identifying helpful online reviews: a product designer's perspective. **Computer-Aided Design**, v. 45, n. 2, p. 180-194, 2013.

LYNTRAS, Miltiadis; POULOUDI, Athanasia. Towards the development of a novel taxonomy of knowledge management systems from a learning perspective: an integrated approach to learning and knowledge infrastructures**. Journal of Knowledge Management**. Vol 10, p. 64-80, 2006.

LONGHI, Magalí T. et al. **Investigando a subjetividade afetiva na comunicação assíncrona de ambientes virtuais de aprendizagem**. XX Simpósio Brasileiro de Informática na Educação, Florianópolis. 2009.

MACDONALD, Craig; et al. Overview of the TREC 2007 Blog Track. In: **TREC**. 2007. p. 31-43.

MALOUF, R.; MULLEN, T. **Taking sides: user classification for informal online political discourse**. Internet Research v. 18, n. 2, p. 177-190, 2008.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to Information Retrieval**. Cambridge University Press. 2009.

_____.; SCHUTZE, Hinrich. **Foundations of Statistical Natural Language Processing**. Cambridge, Massachusetts: MIT Press, 1999.

MARTINEZ, Maria Laura; FERREIRA, Sérgio Leal. **Da Web 2.0 ao Learning 2.0: Novas oportunidades e desafios para o design de interfaces de aprendizagem** (2007). Disponível em: http://www.degraf.ufpr.br/artigos_graphica/DAWEB.pdf. Acesso em: 20 jun. 2014.

MARTÍNEZ-CÁMARA, Eugenio et al. Integrating Spanish lexical resources by meta-classifiers for polarity classification. **Journal of Information Science**, p. 0165551514535710, 2014.

MCNEMAR, Quinn. **Note on the sampling error of the difference between correlated proportions or percentages**. Psychometrika, v. 12, n. 2, p. 153-157, 1947.

MELGAR SASIETA, Héctor Andrés. **Um Modelo para a visualização de conhecimento baseado em imagens semânticas**. (Tese) Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. Universidade Federal de Santa Catarina. Florianópolis, 2011.

MEJOVA, Yelena. **Sentiment Analysis: An Overview**. Comprehensive exam paper, disponível em: http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf. 2011.

MEDHAT, Walaa; HASSAN, Ahmed; KORASHY, Hoda. Sentiment analysis algorithms and applications: A survey. **Ain Shams Engineering Journal**, 2014.

MICHIE, Donald; SPIEGELHALTER, David J.; TAYLOR, Charles C. **Machine learning, neural and statistical classification**. 1994.

MIHALCEA, Rada; STRAPPARAVA, Carlo. Learning to laugh (automatically): Computational models for humor recognition. **Computational Intelligence**, v. 22, n. 2, p. 126-142, 2006.

MINHAS, Saliha et al. A review of artificial intelligence and biologically inspired computational approaches to solving issues in narrative financial disclosure. In: **Advances in Brain Inspired Cognitive Systems**. Springer Berlin Heidelberg, 2013. p. 317-327.

MOLINA-GONZÁLEZ, M. Dolores et al. Semantic orientation for polarity classification in Spanish reviews. **Expert Systems with Applications**, v. 40, n. 18, p. 7250-7257, 2013.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. Conceitos sobre aprendizado de máquina. **Sistemas Inteligentes-Fundamentos e Aplicações**, v. 1, p. 1, 2003.

MORAES, R.; VALIATI, J. F.; GAVIÃO NETO, W. P. Document level sentiment classification: An empirical comparison between SVM and ANN. **Expert Systems with Applications**, v. 40, n. 2, p. 621-633, 2013.

MOREO, Alejandro et al. Lexicon-based comments-oriented news sentiment analyzer system. **Expert Systems with Applications**, v. 39, n. 10, p. 9166-9180, 2012.

MULLEN, Tony; COLLIER, Nigel. Sentiment analysis using support vector machines with diverse information sources. In Proceedings of **Conference on Empirical Methods in Natural Language Processing**, Barcelona, 2004.

NARAYANAN, R.; LIU, B.; CHOUDHARY, A. **Sentiment Analysis of Conditional Sentences**. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 09) v.1, p. 180-189, 2009

NARR, Sascha; HULFENHAUS, M. ALBAYRAK, Sahin. **Language-Independent Twitter Sentiment Analysis**. Processing Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012). Alemanha, 2012.

NASSIRTOUSSI, ArmanKhadjeh et al. Text Mining for Market Prediction: A Systematic Review. **Expert Systems with Applications**, no prelo, 2014.

NIU, Li; LU, Jie; ZHANG, Guangquan. **Cognition in Business Decision Support Systems**. Springer Berlin Heidelberg, 2009.

NONAKA, Ikujiro; KROGH, Georg von. **Tacit Knowledge and Knowledge Conversion: Controversy and Advancement in Organizational Knowledge Creation Theory**. Organization Science. Vol 20, p.635-654, 2009.

O'REILLY, Tim. **What is web 2.0. Design patterns and business models for the next generation of software. 2005**. Disponível em: http://facweb.cti.depaul.edu/jnowotarski/se425/What%20Is%20Web%202%20point%200.pdf . Acessadoem: 20 jun. 2014.

O'REILLY, Tim, **What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software**. Communications & Strategies, No. 1, p. 17, 2007.

OHANA, Bruno; DELANY, Sarah Jane; TIERNEY, Brendan. A case-based approach to cross domain sentiment classification. In: **Case-Based Reasoning Research and Development**. Springer Berlin Heidelberg, 2012. p. 284-296.

PADMAJA, S.; FATIMA, S. S. **Opinion Mining and Sentiment Analysis – An Assessment of Peoples' Belief: A Survey**. International Journal of Ad hoc, Sensor & Ubiquitous Computing (IJASUC) Vol.4, No.1, 2013.

PAI, Mao-Yuan, et al. **Electronic word of mouth analysis for service experience**. Expert Systems with Applications. v.40, p. 1993–2006, 2013.

PAK, Alexander; PAROUBEK, Patrick**. Twitter as a Corpus for Sentiment Analysis and Opinion Mining**. Proceedings of the Seventh conference on International Language Resources and Evaluation LREC'10, Valletta, Malta, 2010.

PANG, Bo; LEE, Lillian. **Opinion Mining and Sentiment Analysis**. Foundations and Trends in Information Retrieval. Vol 2, p. 1-135, 2008.

PARK, Se Jung et al. **Networked politics on Cyworld: The text and sentiment of Korean political profiles**. Social Science Computer Review, v. 29, n. 3, p. 288-299, 2011.

PARK, Sungrae; LEE, Wonsung; MOON, Il-Chul. Efficient extraction of domain specific sentiment lexicon with active learning. **Pattern Recognition Letters**, v. 56, p. 38-44, 2015.

PEÑALVER-MARTINEZ, Isidro et al. Feature-based opinion mining through ontologies. **Expert Systems with Applications**, v. 41, n. 13, p. 5995-6008, 2014.

POLI, Roberto; OBRST, Leo. The interplay between ontology as categorial analysis and ontology as technology. In: **Theory and applications of ontology: Computer applications**. Springer Netherlands, 2010. p. 1-26.

PRABOWO, Rudy; THELWALL, Mike. **Sentiment analysis: A combined approach**. Journal of Informetrics, v. 3, n. 2, p. 143-157, 2009.

QI, Guilin; HARTH, Andreas. **Reasoning with Networked Ontologies**. In: Ontology Engineering in a Networked World. Springer Berlin Heidelberg. p. 363-380, 2012.

QIU, Guanget et al. **Expanding domain sentiment lexicon through double propagation**. Conference on Artificial Intelligence, IJCAI-09, p.1199-1204, 2009.

QIU, Guanget et al. **Opinion Word Expansionand Target Extractionthrough Double Propagation**. ComputationalLinguistics, Vol. 37, No. 1: 9.27, Março, 2011.

RAHAYU, D. A. et al. RnR: Extracting Rationale from Online Reviews and Ratings. In: **Data Mining Workshops (ICDMW)**, 2010 IEEE International Conference on. IEEE, 2010a. p. 358-368.

RAHAYU, Dwi AP et al. Web services for analysing and summarising online opinions and reviews. In: **Towards a Service-Based Internet**. Springer Berlin Heidelberg, 2010b. p. 136-149.

RAO, Yanghui et al. Building emotional dictionary for sentiment analysis of online news. **World Wide Web**, v. 17, n. 4, p. 723-742, 2014.

RAJPER, A. M. et al. **Sentiment Analysis of Enterprise Mashups Using Scikit and NLTK**. Sinth University Research Journal (Science Series). v. 44, n.4, p. 601-604, 2012.

RAUTENBERG, Sandro et al. ontoKEM: uma ferramenta para construção e documentação de ontologias. **Seminário de Pesquisa em Ontologia no Brasil**, v. 1, 2008.

RILOFF, E. et al. **Learning Subjective Nouns Using Extraction Pattern Bootstrapping**. Proceedings of the CoNLL-03 conference. 2003.

RISH, Irina. An empirical study of the naive Bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. 2001. p. 41-46.

ROBALDO, Livio; DI CARO, Luigi. Opinion Mining-ML. **Computer Standards & Interfaces**, v. 35, n. 5, p. 454-469, 2013.

ROTH-BERGHOFER, Thomas; ADRIAN, Benjamin. From provenance-awareness to explanation awareness—when linked data is used for case acquisition from texts. In: **ICCBR 2010 Workshop Proceedings**, Viale Teresa Michel. 2010. p. 103-106.

SALM JUNIOR, J. F. **Padrão de projeto de ontologias para inclusão de referências do novo serviço público em plataformas de governo aberto**. (Tese) Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. Universidade Federal de Santa Catarina. Florianópolis, 2012.

SAM, K. M.; CHATWIN, C. R. **Ontology-Based Sentiment Analysis Model of Customer Reviews for Electronic Product**. In: International Journal of e-Education, e-Business, e-Management and e-Learning, v. 3, n. 6, p. 477-482, 2013.

SANI, Sadiq et al. Should Term-Relatedness Be Used in Text Representation?. In: **Case-Based Reasoning Research and Development**. Springer Berlin Heidelberg, 2013. p. 285-298.

SAKTHIVEL, M.; HEMA, G. **Sentiment Analysis Based Approaches for Understanding User Context in Web Content**. In: International Journal of Computer Science and Mobile Computing, IJCSMC, v. 2, n. 7, p.231-239, 2013.

SCHLEGEL, Daniel R.; SHAPIRO, Stuart C. Inference graphs: A new kind of hybrid reasoning system. In: **Proc. AAAI**. 2014. p. 1.

SCHREIBER, G. et al. **Knowledge engineering and management: the commonKADS methodology**. MIT Press: Cambridge, 2002.

SCHUMAKER, Robert P. et al. **Evaluating sentiment in financial news articles**. Decision Support Systems, v. 53, n. 3, p. 458-464, 2012.

SERRANO-GUERRERO, Jesus et al. Sentiment analysis: A review and comparative analysis of web services. **Information Sciences**, v. 311, p. 18-38, 2015.

SHANKAR, K. B.; KUMAR, S. **Social Media Analytics – Deep Insight**. In: International Journal of Emerging Technology and Advanced Engineering. v. 3, 2013.

SILVA, E. L. da; MENEZES, E. M. **Metodologia da pesquisa e elaboração de dissertação**. 3. ed. rev. atual. Florianópolis: Laboratório de Ensino a Distância da UFSC, 2001.

SIMRANJIT, S.; NIKUNJ, M.; NISHANT, M. **A Multifaceted Approach Towards Friend Recommendation in Social Network**. International Journal of Computer Science and Telecommunications. V. 3, 2012.

SMALL, Henry. **Interpreting maps of science using citation context sentiments: a preliminary investigation**. Scientometrics, v. 87, n. 2, p. 373-388, 2011.

STUDER, Rudi; BENJAMINS, V. Richard; FENSEL, Dieter. **Knowledge engineering: principles and methods**. IEEE Transactions on Data and Knowledge Engineering, 1998.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao Data Mining.** Editora Ciência Moderna, Rio de Janeiro, 2009.

TAN, Songbo; ZHANG, Jin. An empirical study of sentiment analysis for chinese documents. **Expert Systems with Applications**, v. 34, n. 4, p. 2622-2629, 2008.

TANG, Huifeng; TAN, Songbo; CHENG, Xueqi. A survey on sentiment detection of reviews. **Expert Systems with Applications**, v. 36, n. 7, p. 10760-10773, 2010.

TELLES, André. **A Revolução das Mídias Sociais**. São Paulo: M.Books do Brasil Editora Ltda, 2011.

TORRES, Cláudio. **A Biblio do Marketing Digital**. São Paulo: NovatecEditora, 2009.

THET, TunThura; NA, Jin-Cheon; KHOO, Christopher SG. **Aspect-based sentiment analysis of movie reviews on discussion boards**. Journal of Information Science, v. 36, n. 6, p. 823-848, 2010.

THUMS, Jorge. **Educação dos Sentimentos**. Porto Alegre: Editora da Ulbra e Editora Sulina, 1999.

TSYTSARAU, Mikalai; PALPANAS, Themis. **Survey on mining subjective data on the web**. Data Mining and Knowledge Discovery, v. 24, n. 3, p. 478-514, 2012.

TUMASJAN, Andranik et al. **Election Forecasts With Twitter How 140 Characters Reflect the Political Landscape**. Social Science Computer Review, v. 29, n. 4, p. 402-418, 2011.

TURNEY, Peter D. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, Proceedings in: **12th European Conference on Machine Learning**, p.491-502, September 05-07, 2001.

TURNEY, Peter D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of **the 40th annual meeting on association for computational linguistics**. Association for Computational Linguistics, 2003. p. 417-424.

TURNEY, Peter D.; LITTMAN, Michael L. Measuring praise and criticism: Inference of semantic orientation from association. **ACM Transactions on Information Systems (TOIS)**, v. 21, n. 4, p. 315-346, 2003.

VASCONCELOS, José Braga de.; ROCHA, Álvoro.; KIMBLE, Chris. Sistema de informação de memória organizacional: uma abordagem ontológica para a definição de competências de grupo. **Atas da 4ª Conferência da Associação Portuguesa de Sistemas de Informação**, Porto, Portugal, 2003.

VEERASELVI, S.J.; DEEPA, M. **Survey on Sentiment Analysis and Sentiment Classification**. International Journal of Engineering Research & Technology (IJERT). v. 1, 2013.

VELARDI, Paola et al. Evaluation of OntoLearn, a methodology for automatic learning of domain ontologies. In: BUITELAAR, Paul; CIMIANO. Philipp; MAGNINI, Bernardo (Eds.). **Ontology learning from text: methods, applications and evaluation**. Amsterdam: IOS Press, 2003.

VON WANGENHEIM, Christiane A. Gresse; VON WANGENHEIM, Aldo v Aldo. **Raciocínio baseado em casos**. Editora Manole Ltda, 2003.

VOORHEES, E. M.; BUCKLAND, L. P. **The Sixteenth Text REtrieval Conference Proceedings** (TREC 2007). NIST Special Publication, p. 500-274, 2007.

WANG, P.; DOMENICONI, C. **Building semantic kernels for text classification using wikipedia**. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. p. 713-721, 2008.

WANG, Bo; GUO, Xiaojun. Online recruitment information as an indicator to appraise enterprise performance. **Online Information Review**, v. 36, n. 6, p. 903-918, 2012.

WANG, H.; NIE, X.; LIU, L. **A Fuzzy Domain Sentiment Ontology based Opinion Mining Approach for Chinese Online Product Reviews**. JOURNAL OF COMPUTERS, VOL. 8, NO. 9, SEPTEMBER 2013.

WANG, Wei; XU, Hua; WAN, Wei. Implicit feature identification via hybrid association rule mining. **Expert Systems with Applications**, v. 40, n. 9, p. 3518-3531, 2013.

WEBER, R. O.; ASHLEY, K. D.; BRÜNINGHAUS, S. **Textual case-based reasoning**. The Knowledge Engineering Review, Cambridge University Press, v. 20:3, p. 255–260, 2006.

WEI, Wei; GULLA, Jon Atle. **Sentiment learning on product reviews via sentiment ontology tree**. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. p. 404-413, 2010.

WEISS, Sholom M. et al. **Text mining predictive methods for analyzing unstructured information**. New York: Springer, 2005.

WESTERSKI, Adam. **Semantic Technologies in Idea Management Systems: A Model for Interoperability, Linking and Filtering**. (Tese) Ingeniaria de Informática. Universidad Politécnica de Madrid, Madri, 2012.

WIEBE, J. **Learning subjective adjectives from corpora**. Proceedings of the National Conveference on Artificial Intelligence. p. 735-741, 2000.

WIEBE, J. et al. **NRRC summer study Jan Wiebe and group** (University of Pittsburgh) on ‘subjective’ statements. 2002.

WIEBE, Janyce et al. **Learning Subjetive Language**. Journal Computational Linguistics. v. 30, p.277-308, Cambridge, 2004.

WOSZEZENKI, Cristiane R. **Modelo de Descoberta de Conhecimento Baseado em Associação Semântica e Temporal entre Elementos Textuais**. Proposta de Tese submetida ao programa de Pós-Graduação em Engenharia e Gestão do Conhecimento como requisite para Qualificação de Doutorado. Florianópolis, 2014.

WU, Y. et al. **OpinionSeer: Interactive Visualization of Hotel Customer Feedback**. IEEE Transactions on Visualization and Computer Graphics, v. 16, n. 6, 2010

WU, Chi-En; TSAI, Richard Tzong-Han. Using relation selection to improve value propagation in a ConceptNet-based sentiment dictionary. **Knowledge-Based Systems**, v. 69, p. 100-107, 2014.

XU, Kaiquan et al. **Mining comparative opinions from customer reviews for Competitive Intelligence**. Decision support systems, v. 50, n. 4, p. 743-754, 2011.

YANG, Christopher C.; DORBIN NG, T. **Analyzing and visualizing web opinion development and social interactions with density-based clustering**. Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, v. 41, n. 6, p. 1144-1155, 2011.

YOU, Qianhaozhe; ZHANG, Yu-Jin. A New Training Principle for Stacked Denoising Autoencoders. In: **Image and Graphics (ICIG), 2013 Seventh International Conference on**. IEEE, 2013. p. 384-389.

ZABIN, J.; JEFFERIES, A. **Social media monitoring and analysis: Generating consumer insights from online conversation**. Aberdeen Group Benchmark Report, January 2008.

ZANGH, Lei; LIU, Bing. **Identifying Noun Product Features that Imply Opinions**.49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Vol 2. 2011a.

ZANGH Lei; LIU, Bing. **Entity Set Expansion in Opinion Documents**. Proceedings of the 22nd ACM conference on Hypertext and Hypermedia. Eindhoven, 2011b.

ZHANG, Wenhao; XU, Hua; WAN, Wei. Weakness Finder: Find product weakness from Chinese reviews by using aspects based sentiment analysis. **Expert Systems with Applications**, v. 39, n. 11, p. 10283-10291, 2012.

ZHAO, Yanyan et al. Aspect-Object Alignment with Integer Linear Programming in Opinion Mining. **Plos ONE**. 10. p.1-18, 2015.

ZHOU, Lina; CHAOVALIT, Pamwadee. **Ontology-Supported Polarity Mining**. Journal of the American Society for Information Science and Technology. n. 59, pp. 98-110, 2008

ZHU, Jianhan; GONÇALVES, Alexandre L.; UREN, Victoria. **Adaptive named entity recognition for social network analysis and domain ontology maintenance**. Tech Report kmi-04-30. Knowledge Media Institute, The Open University, UK, 2005.

APÊNDICE – Produção do autor durante o doutorado

Este apêndice tem como objetivo apresentar os trabalhos e artigos produzidos durante o período de doutorado. A primeira subseção apresenta os artigos completos publicados em periódicos. A subseção seguinte, apresenta o capítulo de livro produzido. Na sequência, são relacionados os artigos apresentados em congressos, as palestras ministradas e, por fim, a publicação do artigo principal da tese.

Artigos completos publicados em periódicos

CECI, F. ; PIETROBON, Ricardo ; Gonçalves, Alexandre Leopoldo. Turning Text into Research Networks: Information Retrieval and Computational Ontologies in the Creation of Scientific Databases. **Plos One**, v. 7, p. e27499, 2012.

PACHECO, R. C. S. ; SELL, Denilson ; STEIL, A. V. ; CECI, F. . A Revista Brasileira de Ciências Ambientais no contexto do Sistema Brasileiro de CT&I. **Revista Brasileira de Ciências Ambientais**, v. 1, p. 75-100, 2012.

CECI, F. ; WOSZEZENKI, C. R. ; GONCALVES, A. L. . O Uso de Anotações Semânticas e Ontologias para a Classificação de Documentos. **International Journal of Knowledge Engineering and Management**, v. 3, p. 1-15, 2014.

ALVAREZ, G. M. ; CECI, F. . Base de Dados Orientada a Grafos: Um Experimento aplicado na Análise Social. **Revista Tecnologia e Sociedade**, v. 11, p. 127-139, 2015

PACHECO, R. C. S. ; SELL, Denilson ; STEIL, A. V. ; CECI, F. ; FERNANDES, V. ; ANDREOLI, C. V. . A Revista Engenharia Sanitária e Ambiental no contexto do Sistema Brasileiro de CTI. **Engenharia Sanitária e Ambiental**, 2015.

Capítulo de livro

BORDIN, A. S. ; CECI, Flavio ; GONCALVES, A. L. ; GAUTHIER, F. A. O. ; PACHECO, R. C. S. . Análise Bibliométrica e Baseada em Descoberta de Conhecimento em Texto da Produção Científica do SIIEPE Região Sul. In: Fernando Alvaro Ostuni Gauthier, Selvino Assmann, Javier Vernal, Silvia Maria Puentes Bentancourt, Micheline

Guerreiro Krause, Maricel Karina López Torres, Fernanda Martinhago, Jair Zandoná, Silvia Regina Pochmann de Quevedo. (Org.). **INTERDISCIPLINARIDADE: teoria e prática.** 1ed.Florianópolis: EGC, 2014, v. 1, p. 367-386.

Trabalhos completos publicados em anais de congressos

GOMES, Vinicius R.; FIDENCIO, Paulo H. G.; CECI, Flavio ; GONCALVES, A. L.. **Recuperação de Informações Textuais e Multimídia Utilizando Expansão de Consulta a Recursos da WEB 2.0**. In: COMPUTER ON THE BEACH, 2012, Florianópolis. COMPUTER ON THE BEACH, 2012. v. 1. p. 41-50.

BORDIN, A. S. ; CECI, F. ; GONCALVES, A. L. ; GAUTHIER, F. A. O. ; PACHECO, R. C. S. . **Análise da Produção Científica do Simpósio Internacional sobre Interdisciplinaridade no Ensino, na Pesquisa e na Extensão - Região Sul**. In: SIIEPE - Simpósio Internacional sobre Interdisciplinaridade no Ensino, na Pesquisa e na Extensão - Região Sul, 2013, Florianópolis. SIIEPE - Simpósio Internacional sobre Interdisciplinaridade no Ensino, na Pesquisa e na Extensão - Região Sul, 2013.

ANDRIANI, M. L. ; CECI, F. ; SELL, Denilson ; TODESCO, J. L. . **Um Experimento Envolvendo a Geração de Mapas de Tópicos Automatizada a partir dos Dados Abertos do Sistema de Convênios (SICONV)**. In: LOD Brasil - Congresso Linked Open Data Brasil, 2014, Florianópolis. Anais do LOD Brasil. Florianópolis: UFSC/EGC, 2014. v. 1. p. 71-84.

Palestras

CECI, F. **Recuperação de Informação - Do mar de informação a busca semântica**. 2012. (Apresentação de Trabalho/Conferência ou palestra).
Referências adicionais: Brasil/Português; Local: Instituto Stela; Cidade: Florianópolis; Evento: Palestras para a Pós-Graduação de Jornalismo da UFSC; Inst. promotora/financiadora: Universidade Federal de Santa Catarina.

MONDO, T. S. ; CECI, F. **O mundo mágico da análise de emoções e sentimentos on line**. 2015. (Apresentação de Trabalho/Conferência ou

palestra). Palavras-chave: Análise de Sentimentos; Indicadores; Turismo.
Referências adicionais: Brasil/Português; Local: Centro Sul; Cidade: Florianópolis; Evento: THOR Turismo & Hotelaria recebe; Inst. promotora/financiadora: ENCATHO&EXPROTEL 2015.

Artigo principal da tese aceito para evento

CECI, F.; WEBER, R. O.; GONÇALVES, A. L.; PACHECO, R. C. S. Adapting Sentiment with Context. In: **Case-Based Reasoning Research and Development**. Springer International Publishing, 2015. Proceeding ICCBR 2015.

GLOSSÁRIO

Árvore de sentimento: estrutura na forma de grafo, em que cada nó é uma instância ou classe de uma ontologia de domínio combinada com termos polares a fim de representar uma classificação já realizada.

Base de casos: é o repositório em que ficam armazenados os casos já classificados juntamente com a sua árvore de sentimento.

Bases de conhecimento: são bases que armazenam o conhecimento de maneira estruturada e que servem de apoio a sistemas baseados em conhecimento.

Casos: estrutura na qual se organiza o conteúdo de um texto bem como a sua árvore de sentimento resultante do processo de classificação.

Classe: conceitos mapeados e definidos em uma ontologia de domínio.

Conhecimento: é a combinação completa de informação, dados e relações que levam os indivíduos à tomada de decisão, ao desenvolvimento de novas informações ou conhecimentos e à realização de tarefas (FIALHO et al., 2006).

Engenharia do Conhecimento: promove o ferramental para sistematizar e apoiar processos da gestão que culminam na concepção de sistemas de conhecimento (SCHREIBER et al., 2002).

Entidade: termo simples ou composto que faz parte de um domínio de aplicação.

Grau de polaridade: valor que pode ser atribuído a uma orientação semântica.

Informação: é o conjunto de dados devidamente processados e compreensíveis, ou seja, a informação é a disposição dos dados de uma forma que apresentem um significado, criando padrões e acionando significados na mente dos indivíduos (FIALHO et al., 2006).

Instância: elemento gerado a partir do conceito de uma ontologia.

Léxico de sentimento: lista de termos previamente classificados como positivos ou negativos.

Ontologias: é uma especificação formal e explícita de troca de conceitos e relações que existem em um domínio e que são compartilhados por uma comunidade (STUDER; BENJAMINS; FENSEL, 1998).

Orientação semântica: pode ser definida como positiva ou negativa; é forma qualitativa para se definir o grau de polaridade.

Polarização: processo de definição da orientação semântica de um texto.

Reconhecimento de entidades: processo que identifica (reconhece) termos simples ou compostos (por exemplo: Universidade Federal de Santa Catarina) em meio a documentos não-estruturados. É uma técnica da área de extração de informação (EI) que tem como função reconhecer entidades em textos de diferentes tipos e de diferentes domínios (ZHU; GONÇALVES; UREN, 2005).

Review: no contexto desta tese são os textos coletados de uma plataforma de comércio eletrônico contendo opiniões sobre um determinado produto.

Significância estatística: medida qualitativa que atesta se duas técnicas ou algoritmos executam de maneira diferente e se o seu resultado não foi gerado pelo acaso.

Subjetividade: característica de um texto não-objetivo que leva em consideração elementos não explícitos.

Termo polar: um termo que possui uma orientação semântica vinculada a ele.

Threshold: valor de corte que define se um caso pode ser recuperado ou não pelo modelo da tese.