Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Análise de Métodos de Extração de Aspectos em Opiniões Regulares

João Paulo Albuquerque Vieira

Teresina-PI, Setembro de 2018

João Paulo Albuquerque Vieira

Análise de Métodos de Extração de Aspectos em Opiniões Regulares

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI, como parte dos requisitos necessários para obtenção do título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Raimundo Santos Moura

Teresina-PI

Setembro de 2018

João Paulo Albuquerque Vieira
Análise de Métodos de Extração de Aspectos em Opiniões Regulares/ João Paulo Albuquerque Vieira. – Teresina-PI, Setembro de 2018-
53 p.

Orientador: Prof. Dr. Raimundo Santos Moura

Tese (Mestrado) – Universidade Federal do Piauí – UFPI
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação, Setembro de 2018.

1. Extração de Aspectos. 2. Processamento de Linguagem Natural. I. Raimundo Santos Moura. II. Universidade Federal do Piauí. III. Ciência da Computação.

CDU 02:141:005.7

“Análise de Métodos de Extração de Aspectos em Opiniões Regulares”

JOÃO PAULO ALBUQUERQUE VIEIRA

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação do Centro de Ciências da Natureza da Universidade Federal do Piauí, como parte integrante dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Aprovada por:

Prof. Raimundo Santos Moura
(Presidente da Banca Examinadora)

Prof. Sílvio Roberto Fernandes de Araújo
(Examinador Externo)

Prof. Pedro de Alcântara dos Santos Neto
(Examinador Interno)

Prof. Ricardo de Andrade Lira Rabêlo
(Examinador Interno)

Teresina, 05 de setembro de 2018

Aos meus avós Maria Aurora Vieira e Raimundo Moacir Vieira.

Agradecimentos

Primeiramente gostaria de agradecer aos meus pais Zenália Vieira Mota e Paulo Albuquerque de Souza e Silva por todo o amor, incentivo e apoio na realização desse sonho. Aos meu tios Zenaide Mota Belém e Antônio Nonato Sousa Belém por todo o amparo oferecido para concretização desse momento. Aos amigos feitos durante a graduação e mestrado que me acompanharam nessa trajetória e que eu vou levar para vida toda. À minha namorada Emanuella Geovana Magalhães de Souza por todo incentivo e compreensão nos momentos mais árduos dessa jornada. E por último, mas não menos importante, ao meu orientador Prof. Dr. Raimundo Santos Moura que exerceu um papel fundamental e de excelência na minha formação, muito obrigado.

"O futuro dependerá daquilo que fazemos no presente."
(Mahatma Gandhi)

Resumo

No contexto do constante crescimento da Web, diversos serviços foram virtualizados, incluindo o surgimento do comércio eletrônico (*e-commerce*). Tanto tradicionalmente quanto por meio de *e-commerce*, as pessoas necessitam comparar produtos e serviços para nortear suas decisões por meio da análise das características desejadas. A mudança que a Web ocasionou foi a exposição de suas opiniões em sites de compra e venda, fóruns na Web, redes sociais ou ainda grupos de discussão, permitindo sua visualização por qualquer pessoa que necessite. Porém, com o crescimento explosivo da Web e da quantidade de dados gerada diariamente, uma análise manual dessas informações tornou-se impossível, tendo promovido o surgimento da área de Mineração de Opiniões. Um sistema de Mineração de Opiniões consiste em identificar, classificar e sumarizar as opiniões em descrições textuais de consumidores sobre produtos ou serviços. Na literatura, existem algumas abordagens utilizadas na identificação de opiniões para extrair a entidade alvo e seus aspectos, a saber: *i*) extração baseada em frequência; *ii*) extração baseada nas relações sintáticas; *iii*) extração usando aprendizado supervisionado; e *iv*) extração usando modelos de tópicos. Este trabalho apresenta uma análise comparativa entre as principais abordagens usadas na tarefa de extração de aspectos em relatos sobre produtos e serviços em sites Web. Nessa Dissertação foram implementadas adaptações de quatro métodos de extração de aspectos e avaliados em dois Corpora distintos, sendo um em português e outro em inglês. Nos experimentos realizados observou-se que o método usando aprendizado supervisionado (redes neurais convolucionais) obteve melhores resultados sobre os demais.

Palavras-chaves: Mineração de Opiniões. Extração de Aspectos. Redes Neurais Convolucionais.

Abstract

In the context of the constant growth of the Web, several services have been virtualized, including the emergence of e-commerce. Both traditionally and through e-commerce, people need to compare products and services to guide their decisions using the analysis of the desired features. The change that the Web caused was the exposure of their opinions on websites of buy and sell, Web forums, social networks or even discussion groups, allowing their visualization by anyone who needs. However, with the explosive growth of the Web and the amount of data generated daily, a manual analysis of this information became impossible, having promoted the emergence ofthe Opinion Mining field. An Opinion Mining system consists of identifying, classifying and summarizing opinions in textual descriptions of consumers about products or services. In the literature, there are some approaches used at the identification ofopinions to extract the target entity and its aspects, namely: i) frequency-based extraction; ii) extraction based on syntactic relations; iii) extraction using supervised learning; and iv) extraction using topic models. This work presents a comparative analysis between the main approaches used at the task of Extraction of Aspects in reports about products and services on web sites. On this dissertation were implemented adaptations of four methods of extraction of aspects and evaluated in two distinct Corpora, one in Portuguese and another in English. On the experiments performed it was observed that the method using supervised learning (convolutional neural networks) obtained better results on the others.

Keywords: Opinion Mining. Aspect Extraction. Convolutional Neural Networks.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

BOW	*Bag of Words*
CBOW	*Continuous Bag of Words*
CNN	*ConvolutionalNeural Network*
CRF	*ConditionalRandom Fields*
HMM	*Hidden Markov Models*
LDA	*Latent Dirichlet Allocation*
NER	*Named-Entity Recognition*
PL	Padrões Linguísticos
PLN	Processamento de Linguagem Natural
pLSA	*Probabilistic Latent Semantic Analysis*
POS	*Part of Speech*
PPGCC	Programa de Pós-Graduação em Ciência da Computação
RNA	Redes Neurais Artificiais
SF	Substantivos Frequentes
SVM	*Support Vector Machine*
TF-IDF	*Term Frequency-Inverse Document Frequency*
UFPI	Universidade Federal do Piauí
XML	eXtensible Markup Language

Sumário

1 Introdução

Neste capítulo são apresentados o contexto do presente trabalho, as motivações da pesquisa, bem como os objetivos buscados por meio da realização desta Dissertação e, por fim, a organização do presente texto.

1.1 Contexto e Motivação

De acordo com os dados do estudo Global Digital Report 2018[1], realizado pelas empresas We Are Social[2] e Hootsuite[3], temos 7,5 bilhões de pessoas no mundo, sendo que mais da metade acessa a Internet e usa aparelho celular. O mesmo estudo aponta para 3,1 bilhões de usuários ativos nas diversas mídias sociais e para o aumento constante do uso de redes sociais via dispositivos móveis. Nos últimos anos o comportamento desses usuários vem mudando, pois além de consumir conteúdos, eles também estão expondo suas opiniões e experiências, seja sobre um produto que adquiriram, um local que visitaram ou um serviço que utilizaram, proporcionando assim uma maior interação.

Um dos aspectos mais importantes da interação entre as pessoas na Internet é a troca de opiniões e experiências, que tem mostrado ser bastante influente no processo de decisão e comportamento de compra dos indivíduos (BICKART; SCHINDLER, 2001). É comum que consumidores busquem a opinião de outros usuários sobre um produto antes de adquiri-lo, bem como buscar recomendações de amigos e parentes sobre serviços. Existem vários locais, como fóruns, blogs, redes sociais, sites de *e-commerce*, entre outros, onde as pessoas escrevem relatos sobre produtos e serviços que ficam disponíveis para outras pessoas visitarem em busca de opiniões (BONCHI et al., 2011; ZÚÑIGA; JUNG; VALENZUELA, 2012; MILNE; WITTEN, 2013). Estes relatos constituem uma importante fonte de informação adicional que frequentemente não está disponível na descrição do produto ou serviço. Além disso, as informações providas por indivíduos na Internet demonstram ser mais confiáveis que as informações fornecidas pelo vendedor (BICKART; SCHINDLER, 2001).

Os relatos postados nas diversas aplicações da Internet são úteis para entender a opinião que as pessoas possuem sobre um objeto ou tema, gerando um *feedback*, uma vez que podem refletir um sentimento em relação a um determinado assunto ou objeto, expressando uma opinião favorável, desfavorável ou neutra. Além disso, os relatos publicados na Internet também são úteis para compreender as necessidades, preferências e interesses das pessoas

[1] https://digitalreport.wearesocial.com/
[2] https://wearesocial.com/
[3] https://hootsuite.com/

(LI et al., 2013).

O *feedback* é uma importante informação para a tomada de decisão dos usuários. Os consumidores podem tirar proveito da experiência de outros usuários para decidir sobre a aquisição de produtos ou serviços. As organizações, que são motivadas a ter conhecimento da opinião do público alvo,podem melhorar os seus produtos e/ou serviços com base nessas informações, uma vez que os relatos oferecem um *feedback* gratuito e espontâneo, admitindo que são escritos sem obrigações e podem exprimir opiniões sem restrições.

Nas atividades comerciais, o *feedback* é o elemento chave para qualquer organização, pois se ela não tem informações para subsidiar as decisões estratégicas, então, certamente, ela se encontra em desvantagem em relação às outras. Ter conhecimento de informações, tais como resgatar o que a imprensa e as mídias sociais falam sobre a organização, permite que uma decisão seja deliberada mais rapidamente,com consistência, objetividade e precisão. Tal conhecimento tem contribuído para o surgimento de novos paradigmas de gestão empresarial e provocado grandes impactos sociais (PEPPARD; WARD, 2016).

Na era digital, também referida como sociedade da informação, os consumidores passam a fazer parte do funcionamento da empresa, na qual a qualidade dos produtos/-serviços e o atendimento aos clientes são de suma importância para a sobrevivência das organizações, que são totalmente dependentes dos seus sistemas e tecnologias de informação. Dessa forma, as opiniões dos consumidores são extremamente relevantes para o sucesso ou falha de um produto/serviço (PEPPARD; WARD, 2016).

A importância da opinião é tão grande que muitas empresas (por exemplo, marketing, relações públicas,pesquisas) têm seu negócio voltado à obtenção deste tipo de informação. Tradicionalmente, respostas a questões sobre a opinião pública envolve técnicas como pesquisa de campo, telefonemas ou questionários escritos. Entretanto, estas técnicas possuem custos, são restritas a um grupo ou amostra, seu retorno é demorado e, muitas vezes, pouco eficaz. Além disso, a latência da opinião também é alta, devido ao longo tempo necessário entre a coleta dos dados brutos, sua análise e disponibilização dos resultados (BECKER; TUMITAN, 2013).

Com a imensa quantidade de dados produzido na Internet – cerca de 2,5 quintilhões de *bytes* de dados criados diariamente[4], pesquisadores da área de Processamento de Linguagem Natural (PLN) têm buscado extrair informações úteis de dados não estruturados, pois cerca de 95% das informações relevantes são originadas de forma não-estruturada, principalmente em textos como e-mail, pesquisas, *posts* em redes sociais e fóruns, entre outros[5].

Este enorme volume de dados possibilita a ampliação das fontes de opinião quanti-

[4] Disponível em www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/

[5] Disponível em www.clarabridge.com/nlp-natural-language-processing/

tativamente e qualitativamente, torna as formas de coleta mais baratas, e reduz o tempo necessário para disponibilização da informação. Desta maneira, as grandes empresas não precisam mais sair em busca da opinião dos seus clientes em dispendiosas pesquisas de campo, gastando recursos e tempo para ter o *feedback* necessário para melhorar os seus investimentos e, consequentemente, aumentar os lucros (LIU, 2010).

Todavia, uma vez que a quantidade de dados disponíveis para análise extrapola a capacidade humana para uma investigação manual, torna-se imprescindível o uso de métodos e ferramentas capazes de processar automaticamente não apenas o conteúdo dos relatos, mas também a opinião e o sentimento que expressam, conseguindo extrair informações realmente úteis sobre elas e, por fim, adquirir novos conhecimentos.

Apesar das áreas da Linguística e PLN estarem consolidadas, com uma história de sucesso que se iniciou em meados da década de 40 (JONES, 1994), pouca pesquisa foi feita sobre as opiniões e sentimentos das pessoas antes do ano 2000. Segundo Liu (2012), isso está relacionado ao fato de que havia outrora poucos documentos de opinião disponíveis digitalmente. A oportunidade de capturar a opinião de um público geral tem levantado o crescente interesse da comunidade científica (por causa dos grandes desafios abertos) e da comunidade de negócios (em razão dos notáveis benefícios para o *marketing* e as previsões do mercado financeiro). Desde o ano 2000, a área de Mineração de Opiniões tem crescido rapidamente para se tornar um dos tópicos de pesquisa mais ativas no PLN e na Mineração de Dados. Sendo também largamente estudada nas ciências de gestão (ARCHAK; GHOSE; IPEIROTIS, 2007; DELLAROCAS; ZHANG; AWAD, 2007; PARK; LEE; HAN, 2007; CHEN; XIE, 2008).

Aplicações em Mineração de Opiniões se espalharam por quase todos os domínios, desde produtos de consumo, assistência médica, turismo, hotelaria e serviços financeiros à eventos sociais e eleições políticas. Atualmente, há um bom número de companhias, de pequena e larga escala, que construíram ou estão construindo suas próprias aplicações de análise de opiniões e sentimentos para esses propósitos, como SAS[6], IBM[7], Oracle[8], SenticNet[9], Luminoso[10], entre outras.

Formalmente, Mineração de Opiniões (do inglês, *Opinion Mining*), também chamada de Análise de Sentimentos (do inglês, *Sentiment Analysis*), é a área de estudo que analisa as opiniões que as pessoas expressam sobre um determinado assunto em textos escritos em linguagem natural. Esta área de pesquisa apresenta um grande espaço de problemas, muitos nomes relacionados (por exemplo, Análise de Opiniões, Extração de Opiniões, Mineração de Sentimentos, Análise de Subjetividade, Análise de Emoções, entre

[6] Disponível em www.sas.com/software/sentiment-analysis.html
[7] Disponível em www.ibm.com/analytics/
[8] Disponível em www.oracle.com/br/applications/customer-experience/social/index.html
[9] Disponível em www.business.sentic.net/
[10] Disponível em www.luminoso.com/products/analytics

outros) e tarefas ligeiramente diferentes.

O termo Análise de Sentimentos é usado quase exclusivamente na indústria, enquanto que Mineração de Opiniões e Análise de Sentimentos são comumente empregadas na academia (LIU,2015). Não surpreendentemente, existe uma confusão entre os profissionais e, até mesmo, pesquisadores sobre a diferença entre *sentimento* e *opinião* e se a área deveria ser chamada de Análise de Sentimentos ou Mineração de Opiniões. No dicionário Merriam-Webster[11], o *sentimento* é definido como "uma atitude, pensamento ou julgamento estimulado pela emoção", enquanto a *opinião* é definida como "uma visão, julgamento ou avaliação formada na mente sobre um assunto particular". A diferença é bastante sutil e cada uma contém alguns elementos da outra. As definições do dicionário indicam que uma opinião está mais relacionada com uma visão concreta da pessoa sobre algo, enquanto um sentimento está mais ligado com a emoção.

Nessa Dissertação, usamos o termo *opinião* como um conceito amplo que abrange sentimentos, avaliações, ou atitudes e informações associadas, tais como o alvo da opinião e a pessoa que detém a opinião, e o termo *sentimento* é usado apenas para indicar a polaridade (*positiva*,*negativa* ou *neutra*) contida na opinião. Dito isso, como o termo Análise de Sentimento está fortemente associado à descoberta da polaridade – pelas definições das palavras *opinião* e *sentimento* – que não é o foco deste trabalho, decidimos usar o termo Mineração de Opiniões para definir essa linha de pesquisa.

De acordo com Tsytsarau e Palpanas (2012) esta linha de pesquisa pode ser estruturada genericamente em três etapas: *i*) identificar as opiniões expressas sobre determinado assunto ou alvo em um conjunto de documentos; *ii*) classificar a orientação ou polaridade desta opinião, isto é, se tende a ser positiva, negativa ou neutra; e *iii*) apresentar os resultados de forma agregada e sumarizada. A Figura 1 mostra uma visão geral das três etapas da Mineração de Opiniões no processo de extração de características de comentários da Internet.

Destaca-se que a extração de aspectos, na etapa de identificação, é uma das principais tarefas de Mineração de Opiniões e está relacionada à descoberta de conhecimentos por meio de técnicas de análise e extração de dados a partir de textos livres, foco principal dessa Dissertação. Esse processo envolve a aplicação de algoritmos computacionais para processar descrições textuais em busca de informações úteis e importantes. Há, na litetatura, algumas abordagens utilizadas que tentam solucionar essa tarefa, a saber: *i*) extração baseada em frequência; *ii*) extração baseada nas relações sintáticas; *iii*) extração usando aprendizado supervisionado; e *iv*) extração usando modelos de tópicos.

[11] Disponível em www.merriam-webster.com/

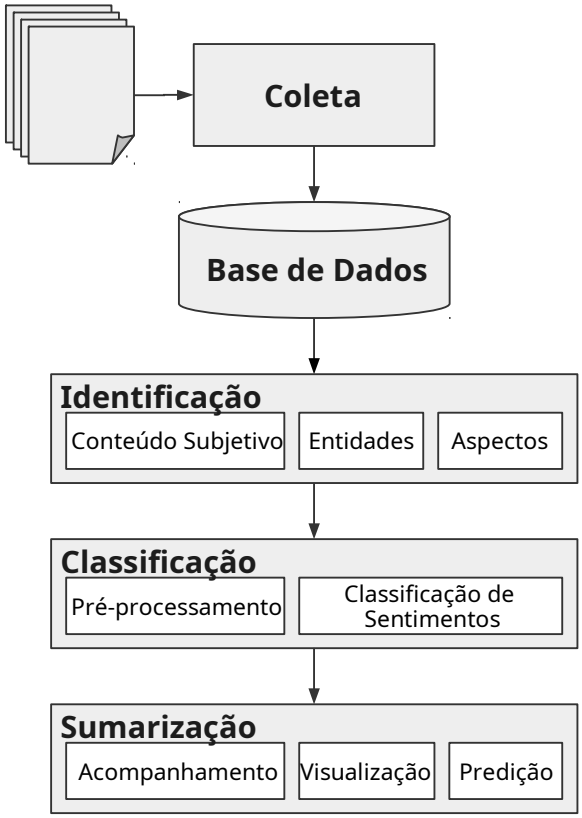

Figura 1 – Etapas da Mineração de Opiniões (BECKER; TUMITAN, 2013).

1.2 Objetivos

Diante da contextualização e dos motivos apresentados, esta Dissertação discute os principais conceitos, técnicas e abordagens computacionais para a Mineração de Opiniões Baseada em Aspectos, mais especificamente para a tarefa de Extração de Aspectos.

O objetivo geral deste trabalho é fazer uma análise comparativa entre as principais abordagens utilizadas para extrair aspectos em textos opinativos no domínio de produtos e serviços. O intuido é avaliar os modelos implementados e descobrir o comportamento das abordagens em *Corpora* do idioma português e inglês. Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Criação de um *Córpus* em língua portuguesa de relatos de consumidores contendo aspectos anotados.
- Implementação de método para extração baseado em frequência, relações sintáticas, aprendizado supervisionado e modelos de tópicos.
- Análise dos métodos implementados e o comportamento entre os *Corpora*.

1.3 Organização

Além deste capítulo introdutório,este trabalho está organizado nos seguintes capítulos:

No Capítulo 2 apresenta o referencial teórico sobre os principais conceitos referentes ao campo da pesquisa que são necessários para o entendimento geral do trabalho.

O Capítulo 3 descreve as principais abordagens existentes relacionadas com a tarefa de Extração de Aspectos em opiniões regulares.

Já no Capítulo 4 são apresentados os experimentos comparativos realizados entre as diferentes abordagens implementadas, além de uma discussão e análise dos resultados.

Finalmente, no Capítulo 5, são apresentadas as considerações finais juntamente com os trabalhos futuros.

2 Referencial Teórico

Neste capítulo serão apresentados os principais conceitos e definições das áreas de PLN e Mineração de Opiniões, com ênfase para a Mineração de Opiniões Baseada em Aspectos. Destaca-se também conceitos da Rede Neural Convolucional (do inglês, *Convolutional Neural Network* - CNN) aplicadas ao processamento de textos e da técnica de modelagem de tópicos conhecida como Alocação Latente de Dirichlet (do inglês, *Latent Dirichlet Allocation* - LDA).

2.1 Processamento de Linguagem Natural

Processamento de Linguagem Natural (PLN), também denominado Linguística Computacional ou, ainda, Processamento de Línguas Naturais, lida com problemas da geração e compreensão automática da língua humana. Os autores Bird, Klein e Loper (2009) definem "linguagem natural" como uma linguagem que é usada para comunicação diária entre humanos como Português, Inglês ou Mandarim. Ao contrário das linguagens artificiais, como linguagens de programação e notações matemáticas, as linguagens naturais têm evoluído à medida que passam de geração para geração, e são difíceis de definir com regras explícitas.

As pesquisas em PLN exploram como os seres humanos entendem e usam a linguagem para que ferramentas e técnicas apropriadas possam ser desenvolvidas. Em termos gerais, tarefas de PLN "quebram" a linguagem em pedaços elementares menores para tentar "entender" os relacionamentos entre eles e explorar como os pedaços trabalham juntos para criar significado.

Algumas técnicas usadas com frequência fazem o uso de conceitos linguísticos como classes gramaticais (substantivos, verbos, adjetivos, advérbios, entre outros), denominadas *Part of Speech* (POS), além de estruturas gramaticais (sintagmas nominais, verbais, preposicionais, entre outros). A área de PLN também lida com situações complexas, como anáforas[1] e ambiguidades[2], bem como o tratamento de erros mecânicos nos textos da *Web*, por exemplo, abreviações, gírias ou neologismos[3]. Isso se dá através de várias representações de conhecimento, como léxicos de palavras e seus significados, propriedades

[1] Anáfora, na linguística, é uma expressão que se refere a uma outra que ocorre na mesma frase ou texto.
[2] Ambiguidade é o nome dado, dentro da linguística, à duplicidade de sentidos, onde alguns termos, expressões, sentenças apresentam mais de uma acepção ou entendimento possível.
[3] Neologismo é um fenômeno linguístico que consiste na criação de uma palavra ou expressão nova, ou na atribuição de um novo sentido a uma palavra já existente.

e regras gramaticais da linguagem, *thesaurus*[4] de sinônimos ou abreviações e ontologias[5] de entidades e ações.

Tecnologias baseadas em PLN estão se tornando cada vez mais difundidas. Ao analisar a linguagem pelo seu significado, os sistemas de PLN podem ser utilizados em diversas aplicações,como por exemplo: *smartphones* que suportam previsão de textos, reconhecimento de voz e correção gramatical; mecanismo de busca *Web* que dão acesso as informações encontradas em textos não-estruturados; e sistemas que permitem tradução automática de textos.

A lista a seguir elenca as principais tarefas e aplicações em PLN:

- Tradução Automática
- Reconhecimento de Entidades Nomeadas
- POS *tagging* (Etiquetagem)
- *Parsing* (Analisadores Sintáticos)
- Perguntas e Respostas
- Mineração de Opiniões
- Reconhecimento de Voz
- Recuperação de Informação
- Extração de Informação
- Sumarização Automática

2.2 Mineração de Opiniões

A Mineração de Opiniões consiste no estudo computacional de opiniões, sentimentos, avaliações, atitudes, e emoções das pessoas através das entidades e de seus diferentes aspectos expressos em textos escritos (LIU,2015). Trata-se de uma área recente que congrega pesquisas de Mineração de Dados, Linguística Computacional, Recuperação de Informações, Inteligencia Artificial, entre outras. O problema de Mineração de Opiniões pode ser estruturado em três tarefas genéricas: *i*) extração de informação; *ii*) classificação da orientação ou polaridade e *iii*) sumarização dos resultados.

[4] *Thesaurus*, também conhecido como dicionário de ideias afins, é uma lista de palavras com significados semelhantes, dentro de um domínio específico de conhecimento.

[5] Ontologia, em ciência da computação, é um modelo de dados que representa um conjunto de conceitos dentro de um domínio e os relacionamentos entre eles.

Segundo Liu (2010), a informação proveniente de um texto pode ser categorizada em dois tipos principais: fatos e opiniões. Um fato é algo que aconteceu na realidade e que é de conhecimento de todos. Uma opinião é uma interpretação dos fatos, que varia de um autor para outro. Liu (2010) ainda define fatos como "expressões objetivas" e opiniões como "expressões subjetivas"Os fatos não podem ser alterados,uma vez que podem ser comprovados através de documentosAs opiniões, por serem subjetivas,divergem de acordo com o sentimento das pessoas que as emitem. Esta característica acrescenta diversos desafios de pesquisa tais como identificação de tópicos e de opiniões. Além disso, elas são escritas em linguagem natural de forma não-estruturada. É necessário, portanto, abstrair conceitos e definições do problema que são abordados a seguir.

2.2.1 Tipos de Opiniões

As opiniões podem ser classificadas em relação a como tratam sobre a entidade alvo e como são expressas no texto. No primeiro caso, são divididas em *regulares* ou *comparativas*. No segundo caso, em *explícitas* ou *implícitas*. Liu (2012) as definem como:

- **Regulares ou Comparativas:** Uma opinião é regular quando o autor da opinião expressa um sentimento, atitude, emoção ou percepção sobre um alvo. As opiniões comparativas expressam o sentimento baseadas na relação de similaridades ou diferenças entre duas ou mais entidades ou quando há algum aspecto compartilhado. Apesar de existir um relacionamento com a opinião direta, estes dois tipos de opiniões apresentam diferenças, tanto semânticas quanto sintáticas. Por exemplo, no relato "*A qualidade da foto da câmera é ótima*" é uma opinião regular – tipo de opinião mais comum, enquanto que no relato "*A qualidade da foto da câmera X é melhor do que a da câmera Y* " é uma opinião comparativa.

  Em destaque, uma opinião regular é definida na literatura apenas como "opinião" e tem mais dois subtipos apresentadas a seguir (LIU, 2007):

  Diretas ou Indiretas: Identifica-se uma opinião direta quando é expressa diretamente sobre a entidade ou um aspecto da entidade – "*a duração da bateria é ótima*", por exemplo, avalia diretamente o aspecto *duração da bateria*. Enquanto que uma opinião indireta é expressa indiretamente, baseando-se nos efeitos que entidade avaliada causa em outra. Esse tipo de opinião é frequentemente encontrado na área médica (LIU, 2012). Como exemplo temos o relato "*após tomar o medicamento, senti fortes dores no estômago*"é possível perceber que não apresenta nenhuma opinião direta sobre a entidade *medicamento*, porém pode-se indiretamente inferir um sentimento negativo devido às dores causadas.

- **Explícitas ou Implícitas:** Nas opiniões explícitas o sentimento está literalmente no texto enquanto que nas opiniões implícitas, geralmente fatos, o sentimento está

expresso indiretamente e tem que ser inferido do textoque às vezes requer um contexto adicional ou conhecimento de domínio. Por exemplo, no relato "*O Android é um sistema operacional ótimo*" é uma opinião explicita pois o sentimento *ótimo* está presente no texto, enquanto que no relato "*comprei o celular há uma semana e já tive que mandar para assistência*" é uma opinião implícita visto que o sentimento não é apresentado textualmente, mas é possível perceber que o celular não atendeu as expectativas do consumidor, expressando um sentimento negativo.

Destaca-se que as opiniões diretas e explícitas são mais fáceis de capturar, concentrando, portanto, a maior parte das pesquisas da área. Opiniões implícitas são difíceis de serem detectadas, pois dependem muito mais do resultado semântico da sentença do que as explícitas e, desse modo, são mais complexas do ponto de vista da PLN. A dificuldade principal deve-se ao fato de que esse tipo de opinião é altamente relacionado ao domínio e ao contexto em que se encontram (ZHANG; LIU, 2011).

Além disso, como as opiniões referem-se a um conteúdo subjetivo,escrito em linguagem natural,a forma como essas opiniões estão expressas tem influência direta na capacidade de processá-las de forma corretaÉ importante destacar que a maioria dos trabalhos concentram-se sobre as opiniões regulares,diretas e explícitas,devido a complexidade na detecção de tais tipos de opiniões (BECKER; TUMITAN, 2013). Nessa Dissertação usamos dois *Corpora* diferentes (mais detalhes na Seção 4.1), que não fazem restrições sobre o tipo de opinião encontrada neles, ou seja, é possível encontrar todos os tipos de opiniões, o que torna a nossa tarefa mais complexa.

2.2.2 Opinião: Definição Formal

Mais formalmente, Liu (2010) define opinião correspondendo a quíntupla $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ onde:

- e_i é o nome de uma entidade;
- a_{ij} é o aspecto da entidade e_i;
- s_{ijkl} é o sentimento sobre o aspecto a_{ij} da entidade e_i, emitido pelo indivíduo h_k no instante t_l;
- h_k é o detentor da opinião (i.e quem emitiu o sentimento);
- t_l é o instante no qual a opinião foi expressa.

Em geral, os cinco componentes são relevantes e não obter um deles ocasiona perda de informação,por exemplo,o aspecto "tela" pode se referir à várias outras entidades

como celulares ou *notebooks*. Em casos específicos, dependendo da aplicação, é possível ignorar componentes que não são considerados relevantes – tendo como exemplo a variável t_l em uma aplicação que não considera o aspecto temporal das opiniões. No entanto, uma opinião é composta por pelo menos dois elementos que são indispensáveis, o alvo da opinião e o sentimento sobre ele (LIU, 2012). O alvo pode ser um aspecto da entidade ou a própria entidade. O sentimento consiste na opinião que o autor da mesma tem sobre o alvo.

Em complemento, o índice subscrito tem como objetivo evidenciar que todos os componentes da quíntupla estão inter-relacionados e que devem ser correspondentes, por exemplo, o sentimento s_{ijkl} deve ser determinado pelo detentor da opinião h_k sobre o aspecto a_{ij} da entidade e_i no instante t_l, no qual qualquer incompatibilidade torna-se um erro. Além disso, também expressa o conceito de que uma entidade pode ter vários aspectos a_{ij}, pois o índice i indica a entidade e o índice j permite uma quantidade qualquer de aspectos. Outra observação é que a definição de Liu (2010) cobre a maioria das opiniões, mas não todos os aspectos do significado semântico da opinião, consistindo numa tarefa complexa. Para exemplificar, o seguinte relato "Este carro é muito pequeno para uma pessoa alta" não diz que o carro é pequeno para todos, sendo que "pessoa alta" é o contexto. Por fim, a definição dada é válida para o tipo de opinião regular, diferente da opinião comparativa que necessita de uma outra definição (JINDAL; LIU, 2006).

As entidades normalmente referem-se a produtos, serviços, organizações, pessoas, eventos, entre outros, e os aspectos são atributos ou componentes das entidades (ZHANG; LIU, 2014). O conceito de aspecto permite que uma entidade seja vista por meio de diferentes perspectivas ou atributos, ou como uma hierarquia de partes e subpartes. Um aparelho celular, por exemplo, é uma entidade, ele possui um conjunto de atributos – como *duração da bateria*, *qualidade da ligação* e *peso* – e um conjunto de partes – como *bateria*, *visor* e *touch screen*. Bateria, por sua vez, também possui um conjunto de atributos, assim como todos os elementos desse grupo composto por diversas partes. Essa representação hierárquica é frequentemente complexa para aplicações. Em problemas práticos, pode ser inviável identificar cada uma das partes da entidade e dos seus respectivos atributos. Desse modo, com o propósito de simplificar a representação apenas dois níveis hierárquicos são considerados e o termo aspecto é utilizado para definir tanto atributos quanto partes constituintes.

Zhang e Liu (2014) além de conceituarem aspecto, também definiram expressão de aspecto como sendo uma palavra ou frase que aparece no texto indicando um aspecto. Como exemplo, os autores utilizam o domínio de celulares e exemplificam o aspecto "qualidade de voz" e discutem que várias expressões como "som" e "voz" podem indicar o aspecto alvo.

É importante observar que os aspectos da entidade são opcionais, pois, às vezes, os usuários podem simplesmente expressar o sentimento sobre a entidade, por exemplo: "*O*

celular é muito bom", nesse caso o sentimento está sendo emitido sobre a entidade *celular*. Em contrapartida, no comentário: "*O display do celular é muito resistente*", o sentimento recai sobre um aspecto do produto, a saber: "*display*". Assim, a entidade e o aspecto juntos representam o alvo da opinião.

A polaridade descreve a direção do sentimento e ela pode ser classificada em classes discretas – *positivo*, *negativo* ou *neutro*, ou como um intervalo que representa a intensidade deste sentimento, tipicamente [-1,1]. Adicionalmente, pode-se apresentar a polaridade em escalas, como *muito positivo* ou *moderadamente positivo*, entre outros (THET; NA; KHOO, 2010).

Outro importante detalhe no que diz respeito as opiniões é que elas podem apresentar diferentes níveis de força (WILSON; WIEBE; HWA, 2004). Por exemplo, no relato "*Este telefone é horrível*" tem força maior do que em "*Este telefone é ruim*". Esta força pode ser interpretada de forma gradativa, por exemplo, uma expressão positiva pode ser expressa através de sentimentos como *contente*, *feliz*, *jubiloso* ou *maravilhado*, partindo de um valor mensurável de baixa intensidade (*contente*) até um valor de máxima intensidade (*maravilhado*) (LIU, 2010).

O detentor da opinião (do inglês, *opinion holder*) é também conhecido como fonte da opinião (KIM; HOVY, 2004; WIEBE; WILSON; CARDIE, 2005). Ele pode ser um indivíduo que emite a opinião de forma pessoal a respeito de uma entidade ou mesmo um porta-voz de uma empresa e tem papel fundamental no processo de avaliar o sentimento de uma determinada opinião. Por fim, o instante no qual a opinião foi expressa é uma importante informação para avaliar o aspecto temporal das opiniões, delimitando como elas mudam com o passar do tempo.

2.2.3 Tarefas da Mineração de Opiniões

Com a definição de opinião bem estabelecida, é possível apresentar as principais tarefas do processo de Mineração de Opiniões. Segundo Liu (2012), o objetivo da Mineração de Opiniões é encontrar todas as quíntuplas $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$ em um documento de opinião. Encontrar cada um dos componentes da tupla pode ser considerado uma tarefa diferente.

A primeira tarefa é encontrar a entidade principal sobre qual a opinião trata. Esse processo se assemelha ao Reconhecimento de Entidades Nomeadas (do inglês, *Named-Entity Recognition* - NER) (MOONEY; BUNESCU, 2005; SARAWAGI, 2008; HOBBS; RILOFF, 2010). NER é uma tarefa de extração de informação que tem como objetivo classificar elementos do texto em categorias pré-estabelecidas, como pessoas, organizações, produtos etc. Em avaliações de consumidores sobre produtos, a entidade geralmente é explícita e pode ser extraída diretamente – pois a avaliação trata de um produto específico. Depois

de extraída, precisamos categorizar as entidades extraídas, pois as pessoas geralmente escrevem a mesma entidade de formas diferentes – por exemplo "*Moto*" e "*Motorola*" que apesar de terem grafias diferentes se referem a mesma entidade.

Por simplificação, os pesquisadores da área com frequência omitem as entidades nas suas discussões e focam apenas nos aspectos. Todavia, frisamos a importância de saber a qual entidade o aspecto pertence, pois é a relação entidade-aspecto que permite identificar apropriadamente as opiniões.

Cada aspecto da entidade também precisa ser identificado. Por exemplo, no relato "*a qualidade das fotos dessa câmera é incrível*", "*qualidade das fotos*" é um aspecto que deve ser extraído. Nesse caso, não é possível utilizar técnicas de NER, pois aspectos não podem ser classificados em categorias pré-definidas pois se apresentam de maneira variada dependendo do contexto que estão inseridosAspectos encontrados em opiniões sobre produtos são diferentes dos encontrados em opiniões sobre política, desse modo, algoritmos de NER não são adequados. Para extração de aspectos são usualmente empregadas técnicas que utilizam a relação semântica entre as palavras que compõe a opinião. Em geral, aspectos são compostos por substantivos e frequentemente estão em frases nominais, mas também podem ser verbos, estar contidos em frases verbais, adjetivos, advérbios e outras construções (HU; LIU, 2004).

Os aspectos ainda podem ser explícitos ou implícitos. Muitos dos aspectos implícitos são adjetivos e advérbios usados para descrever ou qualificar outro aspecto específico, por exemplo, no relato "*Esse celular é muito caro*", a palavra "*caro*" está se referindo ao aspecto *preço*. Eles também podem ser verbos ou frases verbais, por exemplo, no relato "*Esse aparelho pode tocar DVD*", a expressão "*tocar DVD*" caracteriza o aspecto funcional *tocador de DVDs* (HU; LIU, 2004). Já os aspectos implícitos não são apenas adjetivos, advérbios, verbos e frases verbais; eles podem ser arbitrariamente complexos, por exemplo, no relato "*Esse celular pode não caber no meu bolso*", a expressão "*caber no meu bolso*" indica o aspecto *tamanho* (e/ou *formato*), nesses casos algum conhecimento de domínio pode ser necessário para reconhecê-los. A Extração de Aspectos trata-se de um problema desafiador, especialmente quando envolve verbos e frases verbais. Em alguns casos o reconhecimento e anotação dos aspectos torna-se difícil até mesmo para os seres humanos.

A terceira tarefa consiste em identificar o sentimento, polaridade da opinião, normalmente classificado em *positivo*, *negativo* ou *neutro*, sendo que *neutro* significa a ausência de qualquer sentimento expresso ou quando a polaridade dos sentimentos se anulam.

O detentor da opinião refere-se a pessoa ou organização que expressou a opinião – nesse caso, técnicas de NER podem ser utilizadas. Em avaliações de produtos encontradas em sites de compra e venda o autor da postagem geralmente é o detentor da opinião,

podendo ser diretamente extraído. No entanto, em opiniões veiculadas em redes sociais, a extração pode ser de difícil aplicação – pois é comum o compartilhamento de opiniões de terceiros. A extração da informação temporal segue os mesmos princípios, sendo que a data da publicação de uma opinião não é necessariamente a mesma data de sua concepção. Portanto, é preciso definir claramente o significado desse atributo.

Resumindo as descrições acima, dado um conjunto de documentos de opinião d, a Mineração de Opiniões consiste nas seis tarefas enumeradas a seguir:

- **Tarefa 1:** extrair todas as entidades contidas em d e categorizá-las em grupos de entidades sinônimas. Cada grupo representa uma entidade e_i.
- **Tarefa 2:** extrair todos os aspectos das entidades e categorizá-los em grupos. Cada grupo representa um aspecto a_{ij}.
- **Tarefa 3:** extrair todas os detentores das opiniões do texto e categorizá-los de forma análoga às tarefas acima.
- **Tarefa 4:** extrair todas as datas em que as opiniões do texto foram expressas e armazená-las em um formato padrão (por exemplo *dd/mm/aaaa*).
- **Tarefa 5:** para cada aspecto a_{ij}, determinar qual o sentimento relacionado a ele e classificá-lo como *positivo*, *negativo* ou *neutro*, ou ainda o atribuir um valor numérico.
- **Tarefa 6:** gerar todas as quíntuplas de opinião $(e_i, a_j, s_{jkl}, h_k, t_l)$ de cada documento opinativo.

O exemplo abaixo, adaptado de Liu (2012), ilustra o processo da Mineração de Opiniões em relatos da *Web*.

Exemplo: Escrito por: Ricardo Data: 10/06/2017

(1) Comprei uma câmera da Samsung semana passada e meu amigo Pedro comprou uma da Canon. (2) Nos últimos dias, temos usado bastante nossas câmeras. (3) As fotos da minha Samy não são tão boas e duração da bateria é curta. (4) Pedro está bem feliz com a câmera dele e está amando a qualidade das fotos. (5) Quero uma câmera que tire fotos boas também. (6) Vou devolvê-la na loja amanhã.

A **Tarefa 1** deve extrair as entidades "Samsung", "Canon" e "Samy" e agrupar "Samsung" e "Samy", pois representam a mesma entidade. A **Tarefa 2** deve extrair os aspectos "fotos", "qualidade das fotos" e "duração de bateria" e agrupar "fotos" e "qualidade das fotos", pois representam a mesma característica. A **Tarefa 3** deve identificar Ricardo (autor da avaliação) como emissor da opinião da sentença (3) e seu amigo Pedro

como emissor da opinião da sentença (4). A **Tarefa 4** deve encontrar a data "10/06/2017", momento no qual o comentário foi postado. A **Tarefa 5** deve identificar que a sentença (3) expressa opiniões negativas sobre a qualidade das fotos e a duração da bateria da câmera Samsung e que a sentença (4) expressa opinião positiva sobre a câmera Canon de maneira geral e sobre a qualidade das fotos dela. Por fim, a **Tarefa 6** deve gerar as quíntuplas de opinião, que podem ser representadas como:

(Samsung, qualidade_fotos, negativo, Ricardo, 10/06/2017)
(Samsung, duração_bateria, negativo, Ricardo, 10/06/2017)
(Canon, Canon, positivo, Pedro, 10/06/2017)
(Canon, qualidade_fotos, positivo, Pedro, 10/06/2017)

Diante do que foi exposto, consideramos oportuno elencar na próxima seção algumas considerações explicativas sobre os níveis de análise textual.

2.2.4 Níveis de Análise Textual

A detecção de opiniões em um texto pode ocorrer em diferentes granularidades, sendo que a decisão do nível de análise textual está sujeita ao contexto e aplicação. Na taxonomia proposta por Liu (2012), essa análise pode ser efetuada em três níveis, sendo:

- **Nível de Documento:** o objetivo desse nível de análise é classificar a opinião expressa em um documento como um todo. No caso de avaliações de consumidores sobre produtos, esse nível classifica se a avaliação no geral expressa um sentimento negativo, positivo ou neutro sobre o produto. Considera-se que o documento apresenta opiniões referentes apenas a uma entidade (um produto, por exemplo), portanto, esse nível não é adequado para tratar de opiniões comparativas ou que consideram diversas entidades.
- **Nível de Sentença:** o objetivo é determinar o sentimento de uma sentença específica de um certo documento. A utilização mais adequada desse nível é quando um mesmo documento contém opiniões sobre várias entidades. Ele também permite identificar e distinguir sentenças objetivas (fatos) e subjetivas (opiniões).
- **Nível de Entidade-Aspecto:** o nível que realiza a análise com granularidade mais fina, focando na opinião expressa. Em vez de tentar identificar opiniões em estruturas linguísticas (como parágrafos, sentenças e documentos), o nível de entidade-aspecto analisa diretamente as opiniões para então identificar ao que elas estão associadas, ou seja, considera que as opiniões são formadas por alvos e sentimentos. Assim, é possível encontrar em documentos de opiniões diferentes entidades e aspectos dessas entidades, bem como os sentimentos relacionados a eles.

O nível de documento apresenta a análise mais genérica e é o mais limitado, pois avalia a opinião em alto nível para uma entidade desconsiderando seus aspectos individualmente. Além disso, uma opinião pode ao mesmo tempo avaliar diversos aspectos de uma entidade. Assim, é mais adequado para documentos de opiniões concisas, nos quais a entidade avaliada é única e explícita. O nível de sentença é um pouco mais específico e analisa todas as sentenças do documento individualmente. No entanto, apresenta os mesmos problemas do nível de documento, pois considera apenas a sentença em si e não leva em conta a relação entre todas as sentenças que compõe o documento. Por fim, o nível de entidade-aspecto realiza a análise mais detalhada, identificando todas as entidades e os aspectos que são avaliados no documento de opinião, ou seja, cada característica do produto avaliado é considerada importante e deve ser identificada e classificada.

2.3 Mineração de Opiniões Baseada em Aspectos

Como discutido anteriormente, classificar opiniões em nível de documento e em nível de sentença não é suficiente para a maioria das aplicações porque elas não identificam o alvo da opinião, nem atribuem o sentimento para tal alvo. Mesmo quando assumimos que cada documento avalia apenas uma entidade, um documento classificado como uma opinião positiva sobre a entidade não significa que o autor tem opiniões positivas sobre todos os seus aspectos. Da mesma forma, um documento com opinião negativa não significa que o autor seja negativo perante tudo. Para uma análise mais completa, é necessário minerar os diferentes aspectos da entidade e o sentimento expresso em cada um, permitindo detalhar o alvo do sentimento, de tal forma que possam ser detectados seus pontos fortes e fracos.

Em diferentes trabalhos esse nível de análise pode utilizar outros nomes, visto que não existe um termo único que seja natural em todos os domínios. Inicialmente, Hu e Liu (2004) apresentaram o termo Mineração de Opiniões Baseada em Características (do inglês, *Feature-Based Opinion Mining*). Posteriormente o termo Análise de Sentimentos Baseada em Aspectos (do inglês, *Aspect-Based Sentiment Analysis*) foi usado em Liu (2012); contudo em trabalhos onde um tópico denota um aspecto pode ser usado a Análise de Sentimentos Baseada em Tópicos (do inglês, *Topic-Based Sentiment Analysis*). Outros pesquisadores da área utilizam ainda o termo Análise de Sentimentos Baseada em Alvos (do inglês, *Target-Based Sentiment Analysis*) – que inclui tanto as entidades como os aspectos. Nesta Dissertação usamos o termo Mineração de Opiniões Baseada em Aspectos pelo mesmo motivo apresentado na Seção 1.1.

Para alcançar os objetivos da Mineração de Opiniões Baseada em Aspectos é necessário realizar as seis tarefas apresentadas na Seção 2.2.3. Entre essas tarefas, as duas mais importantes são: Extração de Aspectos e Classificação de Sentimentos.

- **Extração de Aspectos:** essa tarefa visa extrair os aspectos que tenham sido

avaliados no texto. Por exemplo, no relato "*a qualidade de vídeo desse celular é decepcionante*",o aspecto "*qualidade de vídeo*" da entidade "*celular* " deve ser extraído. Observando novamente o exemplo anterior, notamos que *celular* não indica um aspecto, visto que a avaliação não é inteiramente referente ao aparelho mas sobre a sua qualidade de vídeo. Contudo, no relato "*Eu amo esse celular* ", onde a opinião refere-se à entidade como um todo,é possível representar o aspecto pelo próprio nome da entidade. Também é comum utilizar o termo "*GERAL*", indicando que a opinião trata sobre a entidade como um todo. Portanto, a Extração de Aspectos também abrange a Extração de Entidades.

A Extração de Entidades e Aspectos são frequentemente consideradas duas tarefas separadas porque os métodos usados para reconhecê-las são normalmente diferentes devido às suas características específicas individuais. As entidades comumente se referem a nomes de produtos, serviços, indivíduos, eventos, e organizações, e os aspectos normalmente se referem aos atributos e componentes das entidades. As duas tarefas, em conjunto, extraem os alvos da opinião. Geralmente,tanto a Extração de Aspectos quanto a Extração de Entidades são tarefas de Extração de Informação. Contudo, no contexto da Mineração de Opiniões, há características específicas do problema que podem facilitar o processo de extração. Uma característica chave das opiniões é que elas sempre tem um alvo – o qual frequentemente indica um aspecto a ser extraídoDesse modo podemos explorar algumas relações das estruturas sintáticas frequentemente usadas para descrever esses relacionamentos entre a opinião e o alvo para ajudar na extração.Em complemento, algumas expressões de opinião podem, além de indicar sentimentos, implicitamente conter aspectos. Por exemplo, no relato "*este celular é caro*" a palavra "*caro*" indica um sentimento negativo, mas também se refere implicitamente ao aspecto *preço* da entidade *celular*, o qual não está presente na opinião. Neste cenário, a utilização de conhecimento de domínio ajuda na identificação dos aspectos (ZHUANG; JING; ZHU, 2006). Após as extrações, uma etapa de agrupamento de entidades e aspectos sinônimos também pode ser aplicada para facilitar a sumarização de opiniões.

A Figura 2 apresenta uma taxonomia das principais abordagens usadas para realizar a análise em nível de entidade-aspecto da tarefa que esta Dissertação aborda.

Exitem quatro abordagens principais para a extração de aspectos explícitos:

1. **Extração baseada em frequência**Utiliza os substantivos e frases nominais mais frequentes encontrados em um grande número de comentários de um dado domínio para identificar aspectos explícitos. Esta abordagem foi desenvolvida inicialmente por (HU; LIU, 2004).

2. **Extração baseada em relações sintáticas**Utiliza analisador gramatical e relações de dependência para obter evidências da relação entre o alvo e as palavras de

sentimento (QIU et al., 2009; ZHANG et al., 2010).

3. **Extração usando aprendizado supervisionado** Utiliza modelos de aprendizado supervisionado para classificação das palavras em *aspecto* e *não-aspecto*. Podemos citar como exemplo os modelos Máquina de Vetor de Suporte (do inglês, *Support Vector Machine* - SVM) (YU et al., 2011), Rede Neural Convolucional (PORIA; CAMBRIA; GELBUKH, 2016), entre outros.

4. **Extração usando modelos de tópicos** Utiliza métodos baseados em agrupamento de tópicos buscando obter distribuições que representem aspectos. Os dois principais modelos encontrados na literatura são Análise Semântica Latente Probabilística (do inglês, *Probabilistic Latent Semantic Analysis* - pLSA) (HOFMANN, 1999) e Alocação Latente de Dirichlet (BLEI; NG; JORDAN, 2003).

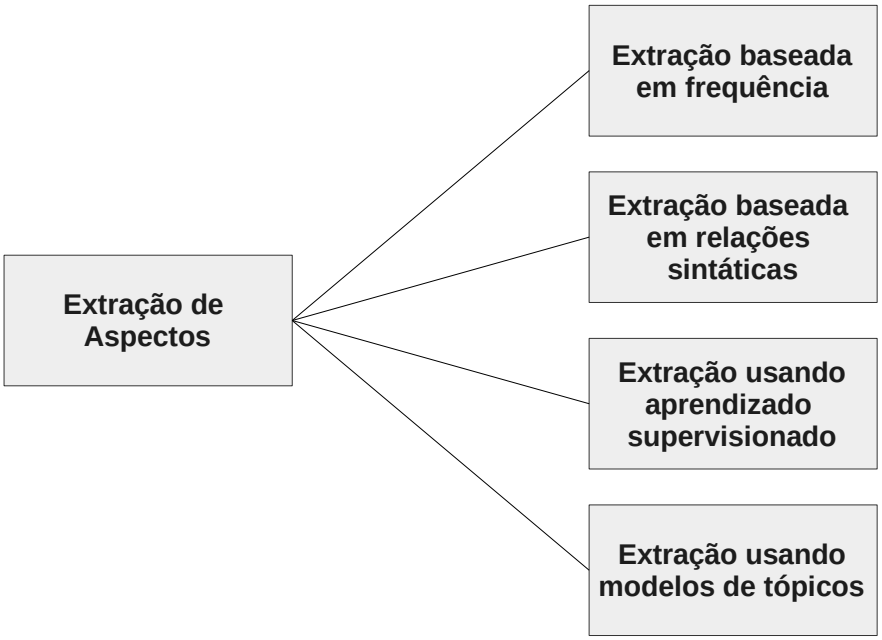

Figura 2 – Taxonomia das principais abordagens usadas para extrair aspectos.

Esta Dissertação foca na extração de aspectos explícitos e suas abordagens serão apresentadas com mais detalhes no próximo capítulo.

2.4 Redes Neurais Convolucionais

As Redes Neurais Convolucionais são Redes Neurais Artificiais em que se aplica a operação de convolução em pelo menos uma de suas camadas. Esse tipo de rede neural foi desenvolvida para um conjunto particular de problemas em que cada amostra segue uma topologia específica. Na topologia, considera-se a existência de uma relação entre valores

de índices próximo em uma representação da amostra. Por exemplo, considerando-se uma imagem de *pixels* como uma matriz,espera-se que valores de índices próximos sejam altamente relacionados. Esse tipo de topologia é explorado em redes convolucionais a partir da sua operação mais fundamental, a convolução.

Define-se a convolução unidimensional discreta entre dois vetores f e g, para x definido no conjunto $\mathbb{Z}$ de inteiros, como:

$$f * g[x] = \sum_{m=-\infty}^{\infty} f[m]g[x-m] \tag{2.1}$$

onde $*$ representa o operador de convolução, m denota o limite inferior e ∞ o limite superior.

Nesse exemplo, fé a entrada, g é o filtro que contém partes dos ganhos da camada e o resultado da convolução entre um dos filtros g e a entrada é chamado de *feature map*. Uma camada convolucional usual contém um conjunto de N filtros g. Como cada filtro g gera, na sua saída, um *feature map*, a saída de uma camada convolucional possui um conjunto de *features maps* $\in R^{N \times M}$, em que N é o número de filtros e M a dimensão do vetor na saída de cada convolução. Assim, no caso particular de uma entrada que f inicialmente era um vetor, a saída de uma camada convolucional passa a ser uma matriz com um número de linhas N igual ao número de filtros g da camada anterior e número de colunas M igual ao tamanho de cada vetor na saída da convolução de cada filtro.

Num caso mais geral em que a entrada f de uma camada já seria uma matriz, a operação que define a convolução unidimensional é modificada. Nessa situação, a convolução entre um filtro g e um conjunto de *features maps* f será a convolução no sentido das linhas da matriz g e do conjunto de *feature maps*, como ilustrado na Figura 3, onde o lado esquerdo ilustra o deslizamento do filtro g sobre a entrada f, o que gera cada elemento do *feature map* no lado direito.

É comum, logo após a convolução, aplicar uma função de ativação. A função de ativação presente em cada neurônio é responsável por aplicar uma transformação nos dados recebidos. Normalmente, utilizam-se funções com algum grau de não-linearidade. A não linearidade das camadas intermediárias permite que as aplicações sucessivas dessas distorções tornem as categorias de saída linearmente separáveis. Alguns exemplos de funções de ativação são: ReLU (NAIR; HINTON, 2010), SELU (KLAMBAUER et al., 2017), ELU (CLEVERT; UNTERTHINER; HOCHREITER, 2015), entre outras.

Um aspecto chave de uma CNN são as camadas de *pooling*, tipicamente aplicadas depois das camadas de convolução. A camada de *pooling* é responsável pela redução da dimensionalidade do modelo. Ao passar uma *feature map* por um *pooling*, definem-se regiões a partir das quais somente um valor será enviado para a camada seguinte. Diferentemente de uma camada de convolução, um *pooling* usualmente não altera o número de *feature*

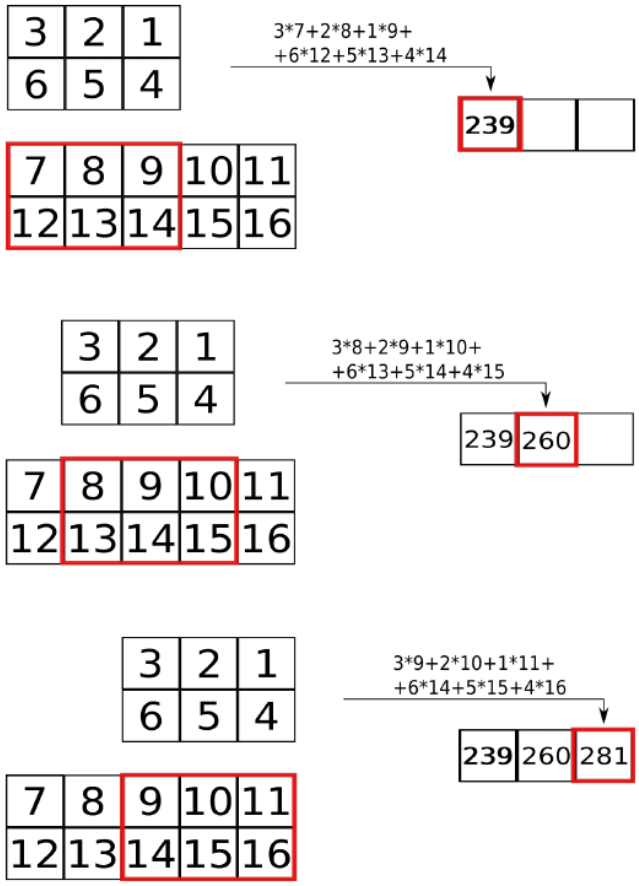

Figura 3 – Convolução unidimensional de um *feature map* 2×5 com um filtro g 2×3.

maps, em vez disso, reduz-se o número de linhas e/ou colunas da entrada. O cálculo da saída de um *pooling* pode ser definido de diversos modos, em que se aplica uma função a todos os elementos de uma mesma região e calcula-se a saída correspondente, alguns exemplos de funções encontradas são *Max Pooling* (o valor da saída será o maior dentre os valores de uma região), *Average Pooling* (o valor da saída será a média de todos os valores da região), entre outras. A forma mais comum é aplicar o *pooling* sobre todo o *feature map* produzindo um único número mas também pode ser aplicado em uma janela, a Figura 4 ilustra a operação de *pooling* sobre filtros em uma matriz. Além de reduzir o tamanho da matriz, consequentemente reduz o processamento para as próximas camadas, essa técnica também auxilia no tratamento de invariâncias locais.

Um dos grandes problemas do treinamento de uma rede é o *overfitting*, ou seja, quando o modelo se torna específico demais para uma determinada base de dados. O *overfitting* pode ser identificado quando há uma grande diferença entre a acurácia do modelo para amostras já observadas e novas amostras. Isso pode ocorrer porque o modelo decora as amostras observadas anteriormente ou também porque ele se ajusta sobre o ruído em vez do verdadeiro modelo subjacente. Uma forma de evitar este problema é a utilização de uma técnica chamada *dropout*, que consiste em desativar um neurônio oculto

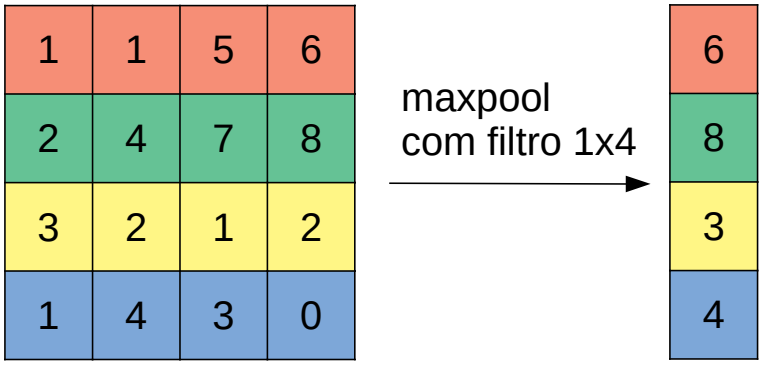

Figura 4 – Aplicação de *Max Pooling* em uma matriz 4×4 utilizando filtro 1× 4.

a partir de uma certa probabilidade. Segundo Krizhevsky, Sutskever e Hinton (2012), esta técnica dá à rede habilidade de aprender características mais robustas, já que um neurônio não pode depender da presença específica de outros neurônios. O resultado da aplicação da técnica pode ser visualizado na Figura 5. Observa-se na figura que os nós marcados com X são desativados, bem como suas ligações.

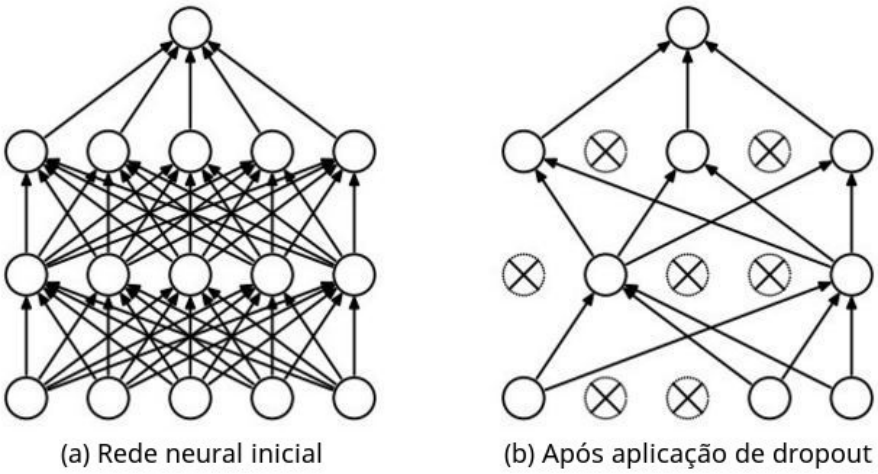

Figura 5 – *Dropout* aplicado em uma rede neural.

O outro tipo de camada encontrada em uma CNN é a completamente conectada. Em arquiteturas de CNNs modernas é comum encontrar ao menos uma camada desse tipo antes da camada de saída. Juntamente com as camadas de convolução e *pooling*, as camadas totalmente conectadas geram descritores de características que podem ser mais facilmente classificados pela camada de saída. Esta camada é responsável por traçar um caminho de decisão a partir das respostas dos filtros vindos das camadas anteriores, para cada classe de resposta.

Depois da camada completamente conectada o último passo é a função de classificação. Essa função realiza a classificação em uma das possíveis classes de saída, trata-se

de uma função fundamental no treinamento, pois influencia no aprendizado dos filtros e consequentemente no resultado da rede.

2.4.1 Redes convolutivas aplicadas no processamento de textos

Originalmente voltado para visão computacional, os modelos de aprendizado profundo têm alcançado notáveis resultados nos últimos anos (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; GRAVES; MOHAMED; HINTON, 2013) e têm subsequentemente mostrado serem eficientes para PLN alcançando bons resultados em *parsing* semânticos (YIH; HE; MEEK, 2014), na recuperação de consultas (SHEN et al.,2014), modelagem de sentenças (KALCHBRENNER; GREFENSTETTE; BLUNSOM, 2014), e outras tarefas tradicionais em PLN (COLLOBERT et al., 2011).

Nesse trabalho usamos uma CNN semelhante ao modelo primeiramente proposto por Kim (2014), representado pela Figura 6, que realiza uma tarefa para a classificação de sentimentos. No entanto, para realizar a tarefa de Extração de Aspectos usamos uma arquitetura ligeiramente diferente, na qual o número de classes de saída correspondente à quantidades de palavras da frase de entrada, ou seja, cada palavra está relacionada com uma classe de saída que pode ser classificada em $aspecto$ ou $não - aspecto$

Para o processamento de textos a entrada para a CNN também é uma matriz, contudo, diferentemente dos modelos utilizados na visão computacional onde essa matriz representa uma imagem, na PLN as linhas das sentenças são representadas como matrizes que servem de entrada para a rede, onde cada linha da matriz corresponde a um *token*, tipicamente uma palavra. Antes do treinamento, as *Word Embeddings*[6] podem ser geradas para cada uma das palavras em um glossário de todas as entradas da sentença. Alternativamente podem ser utilizados modelos já treinados de *Word Embeddings*, por exemplo, *word2vec* (MIKOLOV et al., 2013) ou *GloVe* (PENNINGTON; SOCHER; MANNING, 2014). A sentença tokenizada é convertida em uma matriz de sentença M. As *Word Embeddings* das palavras que correspondem a sentença atual são estão montadas em M, onde as linhas são representações vetoriais das palavras de cada *token*.

Denota-se a dimensionalidade do vetor da palavra por d. Se o comprimento de uma dada sentença é s, então a dimensionalidade da matriz da sentença é $s \times d$. As sentenças que são maiores que o parâmetro responsável por definir o comprimento máximo das sentenças que a rede manipula, são truncadas e as menores são preenchidas com vetores de zero.

Word Embeddings são uma família de técnicas de PLN visando o mapeamento do significado semântico em um espaço geométrico. Isso é feito associando um vetor numérico à cada palavra em um dicionário, tal como a distância (por exemplo, L2 ou mais comumente

[6] *Word Embedding* são representações de palavras em vetores de números reais.

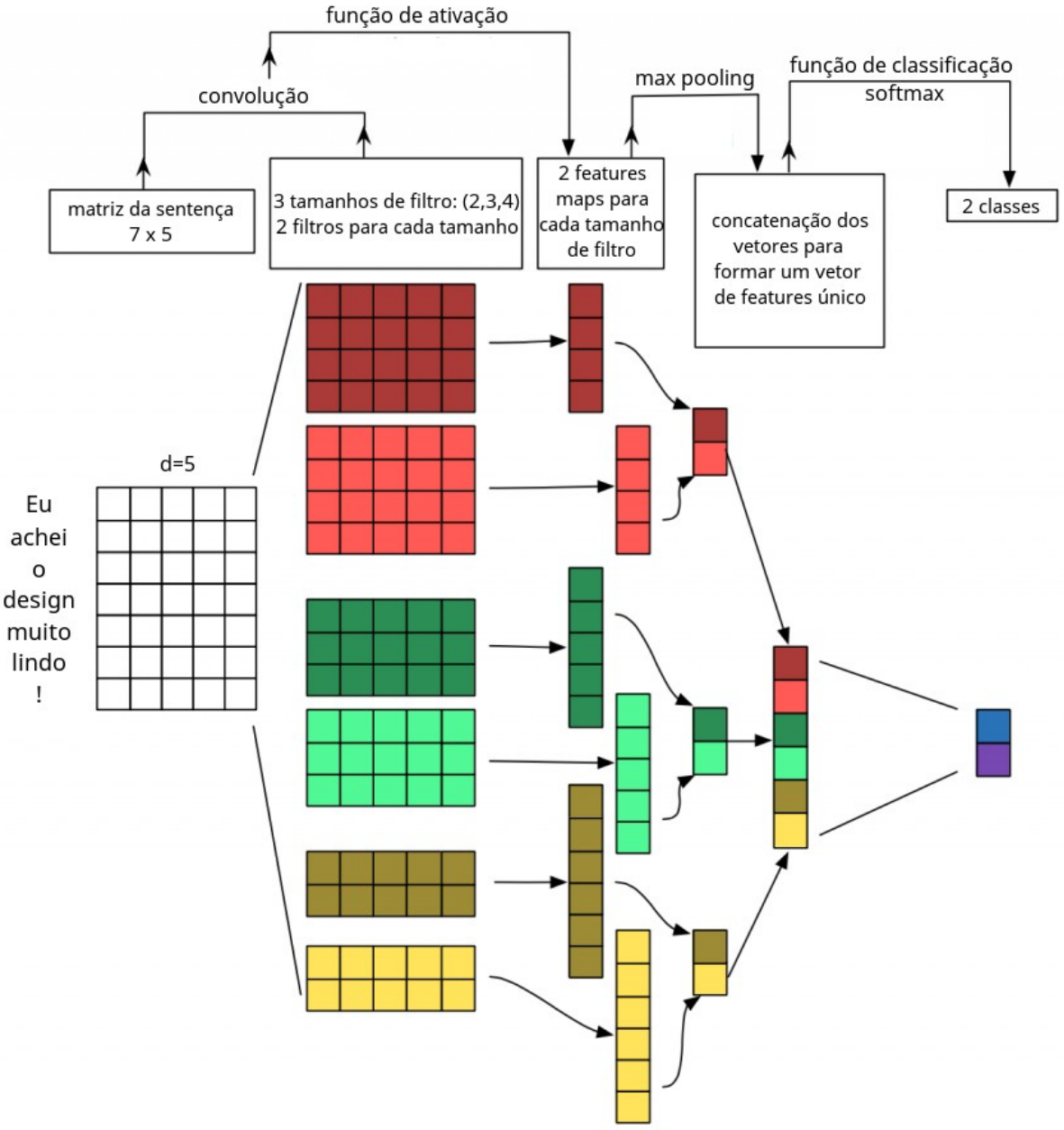

Figura 6 – Modelo CNN para classificação de sentenças (ZHANG; WALLACE, 2015).

cosseno) entre quaisquer dois vetores podendo capturar parte das relações semânticas entre as duas palavras associadas. *Word Embeddings* são calculadas aplicando técnicas de dimensionalidade reduzidas para conjuntos de dados de co-ocorrência estatísticas entre palavras em um *Córpus* de texto. Isso pode ser feito via redes neurais ou matriz de fatoração.

Na camada de convolução, a matriz da sentença é tratada como uma "imagem", e executa-se convoluções sobre ela via filtros lineares. As linhas representam as palavras e normalmente utiliza-se filtros com largura igual a dimensionalidade do vetor de palavras. Assim, podemos simplesmente variar a "altura" do filtro, definido pelo comprimento do

filtro, e executar convoluções de uma dimensão sobre a matriz usando múltiplos filtros com diferentes tamanhos de janela. À medida que os filtros se movem, por várias sequências, eles capturam as características sintáticas e semânticas geradas no *n-grama* filtrado.

Muitas sequências de características resultantes da camada de convolução são combinadas em um *feature map*. Na camada de *pooling*, uma operação de *maxpooling* é aplicada para capturar a característica local mais importante do *feature map* (COLLOBERT et al., 2011). Em seguida, é aplicada *dropout* para controlar o *overfitting* do modelo.

As saídas dos múltiplos filtros são concatenadas para um vetor de características final que é processado por uma camada completamente conectada, cujas saídas passam por uma operação softmax, resultando na distribuição de probabilidades sobre as classes de saída.

2.5 Alocação Latente de Dirichlet (LDA)

LDA é um modelo de aprendizado não-supervisionado que assume que cada documento consiste de uma mistura de tópicos e que cada tópico é uma distribuição probabilística sobre palavras. Trata-se de um modelo generativo[7] de documento que especifica um procedimento probabilístico pelo qual os documentos são gerados. O modelo propõe-se à representar membros de uma coleção de documentos através de descrições sucintas que mantenham as relações estatísticas suficientes para tarefas como classificação de textos, sumarização e recuperação de informação (BLEI; NG; JORDAN, 2003).

Algumas suposições a respeito dos documentos de um *Córpus* e do processo que os gera são feitas pela LDA. Assume-se que os documentos são representados por um modelo *bag of words* (bow[8]) onde a ordem das palavras do documento não é considerada. O processo de elaboração dos documentos considerado no modelo LDA é descrito por Blei, Ng e Jordan (2003) como:

1. Escolhe-se uma quantidade N de palavras.
2. Escolhe-se $\theta \sim Dir(\alpha)$.
3. Para cada uma das N palavras w_n:
   a) Escolhe-se um tópico $z_n \sim Multinomial(\theta)$
   b) Escolhe-se uma palavra w_n a partir de $p(w_n \mid z_n, \beta)$, que é uma probabilidade multinomial condicionada ao tópico z_n

[7] Os modelos generativos descrevem o processo de geração dos dados observados através da especificação de uma distribuição conjunta. Eles contrastam com os modelos discriminativos que modelam diretamente o processo de decisão especificando uma distribuição condicional nos dados observados.

[8] *bow* é uma representação simplificada usada em PLN em que o texto é representado como um saco de suas palavras.

O processo geração de um documento inicia-se escolhendo sua quantidade de palavras (passo 1). A seguir (passo 2) extrai-se um vetor de parâmetros θ a partir de uma distribuição Dirichlet, em que α é um valor de números reais positivos. Dirichlet é uma distribuição sobre distribuições Multinomiais e aqui é utilizada para extrair o vetor de parâmetros θ da distribuição de tópicos utilizada no passo seguinte.

Com isso, para cada palavra escolhe-se um tópico z_n (passo 3.a) a partir de uma distribuição Multinomial com parâmetros θ. Finalmente, escolhe-se uma palavra w_n a partir da probabilidade Multinomial $p(w_n \mid z_n, \beta)$ condicionada ao tópico previamente escolhido (passo 3.b). O parâmetro β desta probabilidade é uma matriz $k \times V$, em que k é a quantidade de tópicos e V é a quantidade de palavras do *Córpus*. Nessa matriz, $\beta(ij) = p(w^j = 1 \mid z^i = 1)$, ou seja, a probabilidade da palavra w^j ser escolhida dado que o tópico z^i foi escolhido.

Para a estimação ou aprendizagem dos parâmetros do modelo LDA existe alguns algoritmos. O algoritmo *Gibbs Sampling* (PORTEOUS et al., 2008) é o mais amplamente utilizado na literatura, relatado como capaz de obter resultados competitivos, com a vantagem na facilidade de entendimento e implementação. Dada uma quantidade k de tópicos que se deseja extrair do conjunto de documentos D, o processo a ser seguido é descrito em alto nível em (BLEI; NG; JORDAN, 2003), como:

1. Percorrer cada documento d em D e atribuir aleatoriamente um tópico z a cada palavra w.

2. Para cada documento d em D:

   a) Calcular a probabilidade $p(z = 1 \mid d = 1)$ para cada tópico z.

   b) Para cada palavra w em d:

      i. Calcular a probabilidade $p(w = 1 \mid z = 1)$.

      ii. Atribuir a w um novo tópico z que maximize a probabilidade $p(z = 1 \mid d = 1) * p(w = 1 \mid z = 1)$

O processo é iniciado atribuindo-se aleatoriamente um tópico z a cada palavra w presente em cada documento do *Córpus* (passo 1). Pode-se notar que com esta atribuição aleatória de tópicos às palavras, já tem-se a representação por tópicos de cada documento, assim como a distribuição de palavras para cada tópico. Contudo, estas representações tendem a ser imprecisas dado que os tópicos foram atribuídos aleatoriamente.

Em seguida, deve-se repetir o passo 2 iterativamente para cada documento da coleção. Neste passo deve-se calcular a probabilidade $p(z = 1 \mid d = 1)$ para cada tópico z, que é a proporção de palavras do documento d atribuídas ao tópico z.

Calcula-se então, para cada palavra w em d, a probabilidade $p(w = 1 \mid z = 1)$, que é a porcentagem de documentos atribuídos ao tópico z e que contém a palavra w. Com estas probabilidades calculadas, deve-se atualizar o tópico atribuído a cada palavra em d, escolhendo-se o novo tópico z que maximize a probabilidade $p(z = 1 \mid d = 1) * p(w = 1 \mid z = 1)$. Após repetir suficientemente o processo acima, as distribuições de probabilidade se estabilizarão.

Com as palavras dos documentos atribuídos aos seus respectivos tópicos, é possível estimar a mistura de tópicos de cada documento e a distribuição de palavras de cada tópico. A mistura de tópicos de um documento pode ser estimada utilizando-se a proporção de palavras deste documento atribuída a um tópico. Já a distribuição de palavras de um tópico pode ser calculada contando-se as quantidades das palavras atribuídas a esse tópico.

2.6 Considerações Finais

Este Capítulo apresentou brevemente a área de Processamento de Linguagem Natural, com ênfase para a tarefa de Mineração de Opiniões Baseada em Aspectos. O minicurso apresentado no Simpósio Brasileiro de Banco de Dados por Becker e Tumitan (2013) é uma boa referência em português sobre o assunto. Na língua inglesa diversos livros têm sido publicados nos últimos anos, com destaque para os livros *Sentiment Analysis and Opinion Mining* por Liu (2012) e *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions* por Liu (2015). Sobre o uso de redes convolucionais aplicadas a tarefas de PLN, o trabalho de Collobert et al. (2011) representa um marco para a área.

3 Abordagens para Extração de Aspectos

Neste capítulo são apresentados os principais trabalhos relacionados à tarefa de Extração de Aspectos, dividindo-os pelas abordagens utilizadas por cada um dos autores. Para essa tarefa existem quatro abordagens principais: extração baseada em frequência; extração baseada em relações sintáticas; extração usando aprendizado supervisionado; e extração usado modelos de tópicos. Em complemento, também apresentamos as linhas de pesquisas desenvolvidas pelo nosso grupo no PPGCC/UFPI.

3.1 Extração baseada em frequência

A extração de aspectos baseada em frequência tem como objetivo encontrar expressões com aspectos explícitos, representados por substantivos e frases nominais a partir de um grande número de comentários em um dado domínio. A abordagem primeiramente identifica os substantivos e frases nominais usando um POS *tagging* e, em seguida, conta suas frequências usando um algoritmo de Mineração de Dados com um limiar de frequência determinado experimentalmente, mantendo-se apenas os substantivos e frases nominais frequentes.

Inicialmente desenvolvida por Hu e Liu (2004a), extração de aspectos baseada em substantivos frequentes funciona porque os possíveis aspectos que são normalmente expressos como substantivos e frases nominais voltam a se repetir por meio de outros usuários que também estão relatando suas opiniões e experiências sobre os diferentes aspectos de uma entidade,os quais acabam normalmente convergindo para o mesmo vocabulário. Desse modo, substantivos e frases nominais que ocorrem com frequência são normalmente aspectos importantes e genuínos.

A abordagem é interessante porque supõe que os usuários tendem a utilizar o mesmo vocabulário para definir o mesmo aspecto, enquanto que o conteúdo irrelevante tende a ser bastante diferente nos diversos comentários, sendo consequentemente infrequente. Por outro lado, os aspectos mais específicos – que não são frequentemente mencionados – não serão encontrados por essa abordagem. Outra suposição feita é que o *Córpus* deve ter um número razoável de comentários e que ele seja sobre um mesmo produto ou pelo menos sobre o mesmo tipo de produto, por exemplo, celulares. Essa abordagem não funciona bem se o *Córpus* tiver um mistura de produtos com grandes diferenças e/ou se cada produto tiver poucos comentários.

Aprimorando essa abordagem, Moghaddam e Ester (2010) estudaram a utilização de coocorrência de palavras buscando identificar os aspectos mais frequentes e remover os

substantivos irrelevantes. A remoção baseou-se no número de ocorrências no *Córpus*, utilizando remoção de afixos e palavras comuns. Scaffidi et al. (2007) propuseram um método que compara a frequência de substantivos e frases nominais do *Córpus* de comentários com as taxas de ocorrências em um outro *Córpus* genérico em inglês para identificar aspectos verdadeiros.

Long, Zhang e Zhu (2010) extraíram os aspectos baseado na frequência e na informações de distância. Primeiramente encontraram o aspecto núcleo usando o método baseado em frequência e depois utilizaram as informações de distância de similaridade usadas em Cilibrasi e Vitányi (2007) para encontrar outras palavras relacionadas ao aspecto.

A extração baseada em substantivos frequentes é a forma mais simples de extrair aspectos e tem servido como *baseline* para métodos mais complexos. Por outro lado, a desvantagem dessa abordagem é que muitos erros de extração ocorrem devido a substantivos frequentes que não representam um aspecto real (QIU et al., 2011).

3.2 Extração baseada em relações sintáticas

A Extração de Aspectos a partir de relações sintáticas foi inicialmente utilizada por (HU; LIU, 2004). Observando a relação entre o alvo e o sentimento, eles utilizaram palavras de sentimento para identificar aspectos poucos frequentes. Tal relacionamento pode ser explorado para extrair entidades e aspectos pois as palavras de sentimento são conhecidas por meio de léxicos de sentimentos por exemplo. Da mesma forma, se alguma palavra de sentimento é desconhecida, essa relação também pode ser usada para extrai-la.

Qiu et al. (2011) propuseram um método para resolver tanto a extração de aspecto quanto a extração de palavras de sentimento. Eles utilizam uma gramática de dependência para descrever as relações sintáticas entre aspectos e palavras de sentimento, considerando que as relações ocorrem somente de forma diretaque aspectos são representados por substantivos e palavras de sentimento por adjetivos.

A abordagem de Sousa, Rabelo e Moura (2015) utilizou a estrutura sintática das sentenças usando POS *tagging* para identificar as características e seus respectivos qualificadores por meio de padrões linguísticos. O autor considerou os verbos como palavras opinativas,além dos adjetivos e advérbios e utilizou padrões linguísticos pré-definidos por Turney (2002) e algumas extensões para satisfazer o domínio de produtos no qual ele trabalhou. De forma semelhante, Liu et al. (2013) utilizam-se destes conceitos para aperfeiçoar o método ao remover palavras que não podem ser aspectos,desta forma, melhorando a precisão do método e sem perda significativa de cobertura.

A principal desvantagem desta abordagem deve-se ao fato de que alguns aspectos

são expressões multi-palavras e muitas vezes estes aspectos estão implícitos. Entretanto, o método é bastante eficaz principalmente por identificar palavras de sentimento independentemente do domínio.

3.3 Extração usando aprendizado supervisionado

A extração usando aprendizagem supervisionada trata a Extração de Aspectos como uma tarefa de classificação. Elas exigem grandes *Corpora* rotulados para treinamento (LIU, 2012). As principais técnicas de Extração de Aspectos com aprendizagem supervisionados utilizam Modelo Escondido de Markov (do inglês, *Hidden Markov Model* - HMM) (RABINER, 1990), Campo Aleatório Condicional (do inglês, *Conditional Random Fields* - CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001) e Rede Neural Convolucional (do inglês, *Convolutional Neural Network* - CNN) (PORIA; CAMBRIA; GELBUKH, 2016).

Jin e Ho (2009) aplicaram um modelo léxico baseado em HMM para identificar padrões e extrair aspectos e expressões de opinião de relatos. Jakob e Gurevych (2010) aplicaram CRF em *Corpora* de diversos domínios. Como atributos, eles utilizaram *tokens*, POS *tagging*, dependência sintáticas, distância entre palavras e palavras de opinião.

Yu et al. (2011) usaram modelos baseados em SVM para classificar os aspectos em um relato. Cada instância representa um aspecto que é classificado como *aspecto* e *não-aspecto*.Todos os aspectos classificados são extraídos e representados através de sumários. O método também identifica expressões e sinônimos de aspectos frequentes. Entretanto, o método apresenta uma limitação importante: a identificação de aspectos só ocorre a partir de uma lista limitada de substantivos frequentes.

Alternativamente, técnicas de aprendizado profundo têm mostrado desempenho promissor em muitas tarefas da PLN, incluindo Mineração de Opiniões (COLLOBERT et al., 2011). Poria, Cambria e Gelbukh (2016) apresentaram a primeira abordagem de aprendizado profundo para a Extração de Aspectos usando uma CNN de sete camadas para etiquetar cada palavra de uma sentença opinativa em *aspecto* ou *não-aspecto*, resultando em uma melhora significativa da acurácia sobre o estado da arte.

3.4 Extração usando modelos de tópicos

A extração usando modelos de tópicos é uma abordagem não-supervisionada que utiliza modelos estatísticos para identificar tópicos em grandes *Corpora*. Quando aplicados ao domínio de relatos de usuários sobre produtos, esses tópicos podem corresponder aos diferentes aspectos, palavras de opinião, ou ainda aos assuntos, especialidades ou locais citados no texto.

Os modelos de extração de tópicos probabilísticos baseiam-se na premissa de que os documentos consistem de uma mistura de tópicose um tópico é uma distribuição probabilística sobre as palavras.Dentre as técnicas encontradas na literatura para a extração de tópicos, as propostas baseadas em extração de dimensões latentes, como pLSA e LDA, se destacaram pela qualidade dos resultados obtidos e pela boa interpretabilidade das dimensões extraídas.

O modelo pLSA introduzido por Hofmann (1999) foio primeiro a formalizar a extração de tópicos probabilísticos. Apesar de prover uma boa base para uma análise dos textos, o modelo pLSA apresenta dois problemas: *i*) o processo de geração dos tópicos para cada documento não é definido, o que exige uma determinada quantidade de parâmetros que cresce linearmente com a quantidade de documentos e que pode levar ao *overfiting* dos parâmetros estimados; e *ii*) o modelo pLSA não determina uma forma natural de calcular as probabilidades relacionadas a um documento que não está no conjunto de treino (BLEI; NG; JORDAN, 2003; KIM et al., 2012). Para evitar esses problemas, os autores Blei, Ng e Jordan (2003) propuseram o modelo LDA.

O modelo LDA é uma extensão do modelo pLSA, que propõe um modelo generativo probabilístico no qual os tópicos são definidos como uma distribuição de probabilidade sobre um vocabulário fixo de termos. Uma característica importante do LDA é que cada relato possui sua própria distribuição de tópicos e, assim, um mesmo relato pode estar relacionado com vários tópicos no qualcada tópico tem sua proporção de relevância.
A distribuição dos tópicos em cada documento obedece à distribuição multivariada de Dirichlet. Esse modelo é uma das técnicas mais proeminentes para a extração de tópicos e se torna atrativo por descobrir grupos de termos que aparecem frequentemente juntos nos documentos, entretanto, os tópicos podem conter aspectos e palavras de opinião e na Mineração de Opiniões estes precisam estar separados (LIU, 2012).

Vários modelos de tópicos, que em sua maioria são extensões da LDA têm sido propostos para realizar a Extração de Aspectos. Branavan et al. (2008) propuseram um método que faz uso das descrições de aspectos encontradas nos campos Prós e Contras dos relatos para ajudar a encontrar aspectos no texto do relato propriamente dito. Seu modelo consiste em duas partes. A primeira parte agrupa as frases-chaves Prós e Contras em categorias de aspectos baseado na distribuição de similaridade. A segunda parte cria um modelo de tópico que modela os aspectos no texto do relato.

Mukherjee e Liu (2012) utilizaram palavras-sementes definidas pelo próprio usuário sobre algum aspecto para produzir distribuições de tópicos. Os autores também apresentam uma estratégia para separar palavras de sentimento e aspectos.

3.5 Trabalhos do grupo de PLN da UFPI

Nesta seção faremos uma breve apresentação sobre os projetos realizados pelo nosso grupo de pesquisa do PPGCC/UFPITais projetos podem fazer uso das técnicas aqui apresentadas para maximizar seus resultados.

Sousa, Rabelo e Moura (2015) propuseram uma abordagem, denominada Top(X) para estimar o grau de importância de comentários sobre produtos e serviços utilizando um Sistema *Fuzzy* com três variáveis de entrada: reputação do autor, quantidade de tuplas <característica, palavra opinativa> e analisador de riqueza e uma variável de saída: grau de importância do comentário.

Santos et al. (2016b) exploraram duas das variáveis de entrada da abordagem Top(X): quantidade de tuplas <característica, palavra opinativa> e analisador de riqueza aplicadas no domínio de hotéis. Além disso, Santos et al. (2016a) também apresentaram abordagens utilizando Sistemas *Fuzzy* e Redes Neurais Artificiais (RNA) a partir das adaptações propostas.

Sá, Vieira e Moura (2017) também utilizaram a abordagem Top(X) como base para propor melhorias, dessa vez, na variável de entrada reputação do autor. Ele fez uma abordagem para definir quais as medidas mais importantes para avaliar a reputação dos autores de comentários *Web* usando uma RNA.

Barbosa, Moura e Santos (2016) fizeram um estudo para examinar quais as variáveis – incluindo padrões linguísticos encontrados na descrição textual – mais importantes para indicar a utilidade de *reviews* de forma automática usando uma RNA. Eles apontam que métodos de extração de aspectos baseados em abordagens sintáticas não são adequados para relatos sobre produtos e serviços.

O que observa-se nos projetos apresentados pelo nosso grupo de pesquisa é que todos eles utilizam uma abordagem pra extrair aspectos baseada em relações sintáticas ao usar os padrões linguísticos encontrados nas descrições textuais. No entanto, não foi realizada nenhuma avaliação para escolher essa abordagem sobre as demais. Portanto, uma análise dos diversos métodos de extração de aspectos deve ser realizada com o objetivo de maximizar o desempenho dos modelos propostos por nosso grupo de pesquisa.

3.6 Considerações Finais

Este capítulo elencou as principais abordagens utilizadas na literatura para a tarefa de Extração de Aspectos na Mineração de Opiniões em nível de entidade-aspecto (ver Figura 2). Adicionalmente, apresentamos uma visão geral dos projetos desenvolvidos pelo grupo de pesquisa em PLN da UFPI. O próximo capítulo apresenta os experimentos realizados e a discussão dos resultados obtidos.

4 Experimentos

De acordo com Pontiki et al. (2014), a Mineração de Opiniões Baseada em Aspectos é analisada na literatura por diferentes tarefas em diferentes *Corpora* de diferentes perspectivas. Essas variações têm dificultado a comparação entre os métodos a serem usados em sistemas de Mineração de Opiniões. Para resolver esse problema, implementamos para cada uma das quatro abordagens apresentada no Capítulo 3 um método correspondente, para então realizarmos uma análise comparativa entre os resultados alcançados pelas as abordagens utilizadas nos métodosAlém disso, comparamos o comportamento das abordagens entre diferentes *Corpora*, sendo um *Córpus* em língua inglesa e o outro em língua portuguesa, no domínio de produtos e serviços.

4.1 *Corpora*

Nesta seção apresentamos com mais detalhes os *Corpora* utilizados nos experimentos para a avaliação dos métodos.

4.1.1 SemEval

Utilizamos o *Córpus* fornecido pela organização do *workshop* SemEval-2014[1], composto por relatos anotados de consumidores no domínio de restaurantes na língua inglesa. Neste *workshop* foram apresentados vários desafios na área de análise semântica computacional, entre eles a Mineração de Opiniões Baseada em Aspectos (Tarefa 4) (PONTIKI et al., 2014). O evento propôs a realização de quatro sub-tarefas, a saber: *i*) identificação de aspectos; *ii*) identificação da polaridade dos aspectos; *iii*) identificação da categoria dos aspectos; e *iv*) identificação da polaridade da categoria dos aspectos. Para esse fim, o evento disponibilizou em formato XML um *Córpus* com anotações manuais dessas informações, um exemplo de relato pode ser visto na Listagem 4.1. A presente Dissertação, em relação ao evento, realizou experimentos sobre a primeira sub-tarefa.

Listagem 4.1 – Exemplo de relato encontrado no *Córpus* SemEval.

```
<sentence id="3359">
    <text>The pizza is the best if you like thin crusted pizza.</text>
    <aspectTerms>
        <aspectTerm term="pizza" polarity="positive" from="4" to="9"/>
        <aspectTerm term="thin crusted pizza" polarity="neutral" from="34"
            to="52"/>
```

[1] Disponível em: alt.qcri.org/semeval2014/task4/index.php?id=data-and-tools

```
    </aspectTerms>
    <aspectCategories>
        <aspectCategory category="food" polarity="positive"/>
    </aspectCategories>
</sentence>
```

Para a avaliação dos resultados a equipe do SemEval-2014 dividiu o *Córpus* em duas partes, como mostrado na Tabela 1, uma para o treinamento e outra para o teste. Esta tabela mostra também informações quantitativas do *Córpus* Buscapé que fora criado com relatos sobre celulares e *smartphones* de usuários do site Buscapé[2] (ver mais detalhes na próxima subseção).

Tabela 1 – Corpora usados nos experimentos.

Córpus	Treino	Teste	Total
SemEval	3.041	800	3.841
Buscapé	23.456	2.606	26.062

4.1.2 Buscapé

Devido aos escassos recursos disponíveis, especialmente para *Córpus* com aspectos anotados em língua portuguesa, esse trabalho criou um *Córpus*, de forma semi-automática, para avaliar os métodos implementados. O *Córpus* criado tem um total de 26.062 relatos com anotações de aspectos, correspondendo a todos os relatos encontrados no site Buscapé na categoria de Celular e *Smartphone* até o dia 15 de novembro de 2017. A anotação foi realizada por três avaliadores[3] da área de PLN e a estratégia utilizada para anotar o *Córpus* é descrita a seguir:

1. Coleta das avaliações dos especialistas (Figura 7) encontradas em cada produto.
2. Anotação manual dos aspectos encontrados nos campos *Prós* e *Contras*.
3. Coleta dos relatos dos consumidores (Figura 8).
4. Anotação automática dos relatos dos consumidores com base nos aspectos extraídos nas avaliações dos especialistas.

Como é necessária uma grande quantidade de relatos, optou-se por realizar uma abordagem semi-automática para a etiquetagem do novo *Córpus*. A estratégia é utilizar os aspectos extraídos manualmente das avaliações de especialistas para anotar automaticamente os relatos dos consumidores. Proveniente da análise manual foram encontrados 106

[2] Disponível em: www.buscape.com.br/celular-e-smartphone
[3] Os avaliadores eram formados por um doutor, um mestre e um mestrando do grupo de PLN da UFPI.

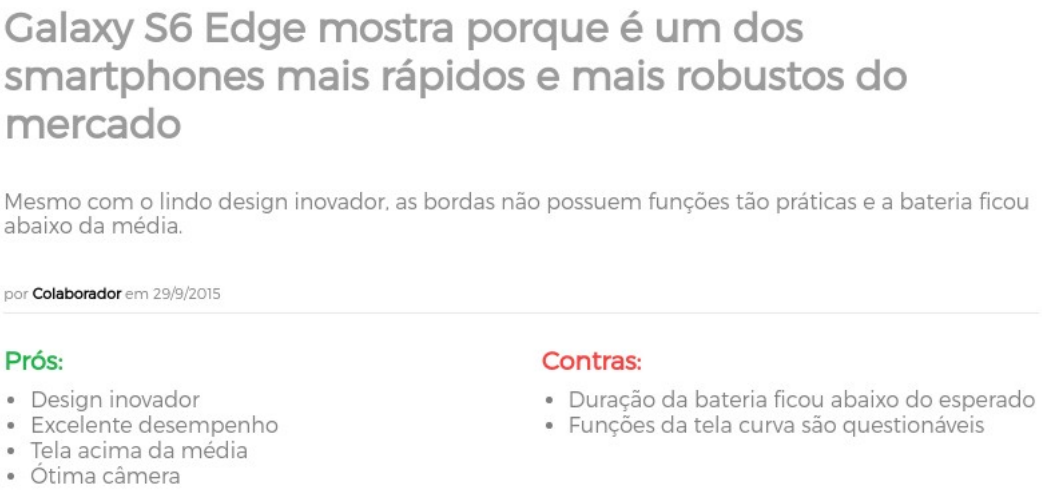

Figura 7 – Avaliação de um especialista sobre o celular Galaxy S6 Edge encontrado no site Buscapé.

Figura 8 – Relato de um consumidor sobre o celular Galaxy S6 Edge encontrado no site Buscapé.

aspectos únicos contidos em um total de 54 avaliações sobre diferentes *smartphones* que foram coletados pela ordem de maior preço na data 11 de janeiro de 2018.

Cada avaliador recebeu uma planilha contendo as sentenças das avaliações dos especialistas encontradas no Buscapé para anotar as palavras que eles consideravam aspectos. Das 54 avaliações coletadas, cada avaliador examinou as mesmas 324 sentenças, encontradas nas *tags* Prós e Contras, em busca de aspectos. As anotações da mesma sentença com mais acertos entre os avaliadores foram definidas como aspectos e, em caso de divergência entre os três, a decisão ficou a critério do mestrando. Os três avaliadores convergiram em 108 das 324 sentenças; tiveram maioria em outras 92 anotações; e os três divergiram em 124 anotações. Isso demonstra que a identificação de aspectos não se trata de uma tarefa trivial. Apesar disso, vale mencionar que a maioria das divergências então relacionadas com a forma como o aspecto foi anotado, por exemplo, na sentença *"O que eu mais gostei foi da câmera traseira dupla"* algumas das formas que os avaliadores poderiam

identificar o aspecto seria: *câmera*, *câmera traseira* ou *câmera traseira dupla*. Ao final obteve-se 84 expressões que representavam aspectos e que seriam usadas para etiquetar os relatos dos consumidores, criando o novo *Córpus*, denominado *Córpus* Buscapé[4].

Uma limitação dessa estratégia de anotação é a possibilidade de existir aspectos que não foram citados nas avaliações dos especialistas, mas que foram citados nos relatos dos consumidores implicando em aspectos não anotados.

Para validação dos experimentos, resolvemos adotar a mesma estratégia usada no *workshop* SemEval-2014 dividindo o *Córpus* em um conjunto de treinamento e um conjunto de validação. Dessa modo, 90% do *Córpus* foi usado para treinamento e os 10% remanescentes para validação, correspondendo a 23.456 e 2.606 relatos respectivamente, ver Tabela 1.

4.2 Métricas de Avaliação

As principais métricas utilizadas para a avaliação dos resultados dos algoritmos de extração de aspectos são os mesmos conjuntos de métricas empregadas nos algoritmos tradicionais de classificação.

- **Acurácia** (A): proporção de classificações corretas, considerando os *aspectos* e *não − aspectos* classificados corretamente sobre a soma de classificações corretas e incorretas;

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{4.1}$$

- **Precisão** (P): percentual de aspectos preditos como sendo realmente um aspecto;

$$P = \frac{VP}{VP + FP} \tag{4.2}$$

- **Revocação**(R): percentual de aspectos que foram corretamente classificados como aspectos;

$$R = \frac{VP}{VP + FN} \tag{4.3}$$

- **Medida-F**(F): As medidas de precisão e revocação podem ser enganosas quando examinadas separadamente. A medida-F busca equilíbrio entre ambas, por meio de uma média harmônica entre elas;

$$F = 2 \times \frac{P \times R}{P + R} \tag{4.4}$$

[4] Disponível para download em: github.com/joaopaulobq/corpusbuscape

onde VP, VN, FP, e FN referem-se a verdadeiro positivo,verdadeiro negativo, falso positivo, e falso negativo, respectivamente. Para uma análise mais detalhada também disponibilizamos as matrizes de confusão dos experimentos realizados sobre cada abordagem que segue o modelo apresentado na Tabela 2. A matriz de confusão é uma tabela que organiza os resultados obtidos pelos algoritmos, permitindo uma comparação da classificação do algoritmo com o resultado esperado.

Tabela 2 – Matriz de Confusão

Valor real	Valor predito	
	aspecto	**não-aspecto**
aspecto	VP	FN
não-aspecto	FP	VN

4.3 Métodos Implementados

Nesta seção descrevemos os métodos que implementamos para cada umas das principais abordagens apresentadas para tarefa de extração de aspectos.

4.3.1 Substantivos Frequentes (SF)

O método usa os substantivos frequentes para extrair os aspectos mais comentados pelos consumidores. Trata-se de um algoritmo simples, onde: após carregar o *Córpus* do banco de dados MySQL, converte-se o texto para minúsculo, transforma-se o texto em *tokens*; elimina-se as contrações (por exemplo, "*da*" para "*de a*" ou "*don't*" para "*do not*"), etiqueta-se cada um dos *tokens* em sua respectiva classe gramatical, extraí-se todos os substantivos e frases nominais usando expressões regulares, aplica-se *stemmer* nas palavras extraídas, e, por fim, aplica-se o algoritmo de Mineração de Dados Apriori (AGRAWAL; SRIKANT, 1994) para a descoberta do conjunto de aspectos frequentes por meio das regras de associações geradas.

Para isso fez-se grande uso da biblioteca NLTK[5] (*Natural Language Toolkit*). O etiquetador *Perceptron Tagger* foi usado para o *Córpus* em inglês e para o português treinou-se um POS *Tagger* a partir do *Córpus* MacMorpho composto por 1,1 milhão de palavras validadas manualmente com anotações morfossintáticas (ALUÍSIO et al., 2003). Para retirar o radical das palavras usamos o Porter *Stemmer* (PORTER, 1980) para o inglês e RSLP *Stemmer* (ORENGO; HUYCK, 2001) para o português.

[5] Disponível em: www.nltk.org

4.3.2 Padrões Linguísticos (PL)

A extração de aspectos baseada em relações sintáticas se fundamenta na premissa que nos relatos de produtos e serviços é comum encontrar características citadas pelos autores nas proximidades das suas respectivas qualidades. E para capturar essas relações implementou-se um método que faz uso de padrões linguísticos. Para o *Córpus* em inglês utilizou-se os padrões especificados por Turney (2002), mostrados na Listagem 4.2. Já no *Córpus* em português uso-se os padrões de Sousa, Rabelo e Moura (2015), mostrados na Listagem 4.3.

Listagem 4.2 – Padrões linguísticos definidos por Turney (2002).

```
1 <ADJ><SUBS>
2 <ADV><ADJ>?!<SUBS>
3 <ADJ><ADJ>?!<SUBS>
4 <SUBS><ADJ>?!<SUBS>
5 <ADV><VERB>
```

Listagem 4.3 – Padrões linguísticos definidos por Sousa, Rabelo e Moura (2015).

```
1 <ADJ><SUBS>(<PREP>?<SUBS>)*
2 <ADV><ADV>?<ADJ>(<SUBS>(<PREP>?<SUBS>)*)?
3 <SUBS>(<PREP>?<SUBS>)*(<ADJ>)<ADV><ADV>?
4 <SUBS>(<PREP>?<SUBS>)*<ADV>?<ADJ>+
5 <ADV><VERB>
6 <VERB><ADV>
```

O método implementado para extrair aspectos baseado na relações sintáticas também fez uso da biblioteca NLTK para o procedimento de tokenização e etiquetagem. Primeiro o texto foi convertido em uma sequencia de *tokens*, processo conhecido como tokenização. Posteriormente etiquetou-se os *tokens* usando um POS TaggerPor fim, extraiu-se dos relatos as palavras cujas *tags* correspondem aos padrões linguísticos definidos por Turney (2002) para relatos em nosso domínio de produtos e serviços em língua inglesa. Para os relatos em língua portuguesa utilizou-se os padrões definidos por Sousa, Rabelo e Moura (2015) que foram inspirados nos padrões de Turney (2002) com adaptações para o português. Nesta tarefa em específico apenas os substantivos são classificados como aspectos apesar da extração ser realizada pelas relações sintáticas entre as palavras.

4.3.3 Rede NeuralConvolucional(CNN)

A CNN foi implementada usando a biblioteca Keras[6] as tem uma arquitetura básica com uma camada de entrada, uma camada de convolução, uma camada de *pooling*, *Dropout*

[6] Disponível em: keras.io

e uma camada totalmente conectada com saída *Softmax*. A estrutura básica foi usada para evitar uma vantagem sobre os outros métodos implementados que também são básicos.

A Tabela 3 apresenta os parâmetros da CNN utilizada nos experimentos. A entrada da rede é uma matriz de *embeddings* com 300 dimensões. Em sua camada de convolução tem-se 100 filtros de comprimento 10. A camada de *pooling* segue a camada de convolução com *max-pooling* de tamanho 2. As saídas das camadas de convoluções foram computadas usando uma função de ativação, a Tabela 4 mostra os resultados de um série de experimentos que foram realizados utilizando diversas funções de ativação disponíveis na biblioteca Keras. Como destacado na tabela, a função Relu apresentou melhor desempenho em todas as medidas analisadas.

Tabela 3 – Valores dos parâmetros usados na CNN.

Nome do parâmetro	Valor
embedding_dim	300
num_filters	100
filter_lenght	10
pool_size	2
batch_size	8
epochs	30
activation_function	*relu*
optimizer	*adam*
loss	*categorical_hinge*

Tabela 4 – Experimento com funções de ativação

Função de ativação	Acurácia	Precisão	Revocação	Medida-F
elu	89.83%	77.40%	23.06%	35.53%
linear	88.39%	55.05%	24.49%	33.90%
relu	**90.67%**	**83.93%**	**28.78%**	**42.86%**
selu	89.78%	76.35%	23.06%	35.42%
softplus	90.38%	81.88%	26.73%	40.31%
softsign	90.33%	79.76%	27.35%	40.73%
tanh	90.25%	80.12%	26.33%	39.63%

No processo de aprendizagem do modelo também realizou-se mais alguns experimentos. Dessa vez sobre a função objetivo que alcançou melhores resultados com a função *categoricalhinge*, como destacado na Tabela 5 e algoritmos de otimização alcançando melhores resultados com o algoritmo Adam (ver Tabela 6).

Tabela 5 – Experimento com funções objetivas.

Função objetivo	Acurácia	Precisão	Revocação	Medida-F
binary_crossentropy	88.59%	61.54%	16.33%	25.81%
categorical_crossentropy	87.38%	45.62%	20.20%	28.01%
categorical_hinge	**90.35%**	**80.98%**	**26.94%**	**40.43%**
cosine_proximity	87.60%	46.09%	12.04%	19.09%
hinge	83.18%	19.08%	11.84%	14.61%
kullback_leibler_divergence	87.52%	46.92%	20.20%	28.25%
logcosh	89.73%	74.68%	23.47%	35.71%
mean_absolute_error	83.18%	19.08%	11.84%	14.61%
mean_squared_error	89.06%	66.23%	20.41%	31.20%
mean_squared_logarithmic_error	89.73%	79.23%	21.02%	33.23%
poisson	87.40%	45.50%	18.57%	26.38%
squared_hinge	87.45%	46.85%	24.29%	31.99%

Tabela 6 – Experimento com algoritmos de otimização

Algoritmo de otimização	Acurácia	Precisão	Revocação	Medida-F
adadelta	88.32%	57.25%	15.31%	24.15%
adagrad	88.54%	59.21%	18.37%	28.04%
adam	**89.09%**	**67.36%**	19.80%	**30.60%**
adamax	88.74%	62.50%	18.37%	28.39%
nadam	87.57%	47.34%	**20.00%**	28.12%
rmsprop	87.08%	42.86%	18.98%	26.31%

4.3.4 Alocação Latente de Dirichlet (LDA)

Utilizou-se uma LDA com algoritmo *Gibbs Sampling* para estimar os parâmetros do modelo e pequenas modificações entre inglês e o português como a lista de stopwords[7], o *Stemmer* e a remoção de contrações.

Um experimento foi realizado para definir a quantidade de tópicos e observou-se que quanto maior o número de tópicos piores são os resultados, por exemplo, com dois tópicos a precisão baixou em 1,34% e a revocação baixou em 12,64%, sempre usando um total de 50 palavras. Vale ressaltar que tópicos diferentes podem conter as mesmas palavras explicando a queda dos resultados. A Tabela 7 apresenta os valores dos parâmetros escolhidos nos demais experimentos.

4.4 Resultados e Discussões

As Tabelas 8 e 9 apresentam as matrizes de confusão geradas com o método Substantivos Frequentes (SF) para os *Corpora* SemEval (inglês) e Buscapé (português),

[7] Em linguagem natural, são palavras vazias, sem valor, normalmente removidas antes ou após o processamento de um texto.

Tabela 7 – Valores dos parâmetros usados na LDA.

Nome do parâmetro	Valor
alpha	0.5
beta	0.5
número de tópicos	1
iterações	50

respectivamente.

Tabela 8 – Matriz de confusão: SF - SemEval.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	1052	602	1654
não-aspecto	1438	7908	9346
Total	2490	8510	11000

Tabela 9 – Matriz de confusão: SF - Buscapé.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	1392	950	2342
não-aspecto	4087	40235	44322
Total	5479	41185	46664

As Tabelas 10 e 11 são as matrizes geradas pelos métodos Padrões Linguísticos (PL) para os mesmos *Corpora*. As Tabelas 12 e 13 referem-se ao método baseado em CNN e as Tabelas 14 e 15 são as matrizes de confusão geras pelo método baseado em LDA.

Tabela 10 – Matriz de confusão: PL - SemEval.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	356	1298	1654
não-aspecto	350	8996	9346
Total	706	10294	11000

Tabela 11 – Matriz de confusão: PL - Buscapé.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	604	1738	2342
não-aspecto	1711	42611	44322
Total	2315	44349	46664

As Tabelas 16 e 17 mostram as métricas de avaliação calculadas para cada um dos algoritmos considerando os *Corpora* SemEval e Buscapé, respectivamente.

Tabela 12 – Matriz de confusão: CNN - SemEval.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	529	1125	1654
não-aspecto	30	9316	9346
Total	559	10441	11000

Tabela 13 – Matriz de confusão: CNN - Buscapé.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	837	1505	2342
não-aspecto	28	44294	44322
Total	865	45799	46664

Tabela 14 – Matriz de confusão: LDA - SemEval.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	481	1173	1654
não-aspecto	1061	8285	9346
Total	1542	9458	11000

Tabela 15 – Matriz de confusão: LDA - Buscapé.

Valor real	Valor predito		Total
	aspecto	**não-aspecto**	
aspecto	1619	723	2342
não-aspecto	8141	36181	44322
Total	9760	36904	46664

Tabela 16 – Comparação entre abordagens usando o *Córpus* SemEval.

Abordagem	Acurácia	Precisão	Revocação	Medida-F
SF	81.45%	42.25%	**63.60%**	**50.77%**
PL	85.02%	50.42%	21.52%	30.17%
CNN	**89.50%**	**94.63%**	31.98%	47.81%
LDA	79.69%	31.19%	29.08%	30.10%

Tabela 17 – Comparação entre abordagens usando o *Córpus* Buscapé.

Abordagem	Acurácia	Precisão	Revocação	Medida-F
SF	89.21%	25.41%	59.44%	35.60%
PL	92.61%	26.09%	25.79%	25.94%
CNN	**96.71%**	**96.76%**	35.74%	**52.20%**
LDA	81.00%	16.59%	**69.13%**	26.76%

Observa-se a baixa precisão do método SF no *Córpus* Buscapé (25,41%) em consequência do grande número de falsos positivos (ver Tabela 9)Decidiu-se, então,

realizar uma análise nas palavras preditas pelo método e verificou-se que muitas das palavras não eram substantivos, ou seja, o etiquetador para português classificou a classe gramatical incorreta, muitas vezes decorrente de erros ortográficos do texto. Portanto, o método SF é totalmente dependente da forma como o *Córpus* está escrito e dos recursos computacionais utilizados, como etiquetadores e *stemming*. Já para o *Córpus* SemEval, obteve-se uma precisão melhor (42,15%) decorrente da escassez de erros ortográficos. Com relação à revocação o método SF alcançou o melhor resultado do *Córpus* SemEval (63,60%) e o segundo melhor resultado do *Córpus* Buscapé (59,44%), comprovando que os substantivos frequentes são verdadeiros indicadores de aspectos.

Como o método PL se baseia nas classes gramaticais das palavras para formar os padrões linguísticos a identificação incorreta das *tags* pelo etiquetador também prejudicou bastante a precisão do método, como pode ser observado nas Tabelas 16 e 17, a saber: 50,42% para o *Córpus* SemEval e 26,09% para o *Córpus* Buscapé. Isso também implica que esse método é totalmente dependente da qualidade de escrita do texto. Curiosamente, Sousa, Rabelo e Moura (2015) também usaram um *Córpus* coletado do Buscapé no domínio de celulares e os padrões sugeridos por eles não obtiveram bons resultados comparado com os padrões de Turney (2002) para a língua inglesa, apesar de serem semelhantes. Alguns fatores foram levantados para justificar essa incongruência: *i*) o etiquetador para a língua inglesa é superior ao etiquetador da língua portuguesa; *ii*) os padrões linguísticos definidos por Turney (2002) conseguiram extrair mais aspectos no *Córpus* SemEval; e *iii*) o *Córpus* Buscapé contém muitas palavras com erros ortográficos, prejudicando a etiquetagem. Por consequência, o método PL obteve uma baixa revocação pois foram identificados poucos padrões linguísticos, fazendo o método "chutar" menos.

A abordagem supervisionada CNN obteve resultados semelhantes nos dois *Corpora*, conseguindo uma precisão de 94,63% no *Córpus* SemEval e 96,67% no *Córpus* Buscapé. No entanto, apresentou baixa revocação o que prejudicou a Medida-F. Isso se deve porque as classes são bastantes desbalanceadas, apenas 15% e 5% do total de palavras são *aspectos* e todo o restante *não-aspectos*, correspondendo respectivamente aos *Corpora* SemEval e Buscapé.

O modelo LDA foi o que apresentou maior divergência entre os *Corpora*. Aplicado ao *Córpus* Buscapé ele encontrou a maior revocação dentre os métodos analisados (69,13%), contudo, apresentou também a pior precisão (16,59%). Já no *Córpus* SemEval obteve os piores resultados, como a saída da LDA é formada por *unigramas* (palavras únicas) e as expressões que representam aspectos do *Córpus* SemEval possuem estruturas multi-palavras bem mais complexas (como as mostradas em Listagem 4.1) isso acabou acarretando nos baixos resultados obtidos. Além disso, uma análise das palavras preditas como *aspectos* pelo modelo LDA também foi feita e observou-se que as palavras selecionadas no modelo de tópicos incluíam palavras de sentimentos, entidades, entre outros ruídos.

Olhando para acurácia que leva em consideração também os *não-aspectos* corretamente classificados, diferentemente das outras métricas, destacamos que o método de aprendizado supervisionado (CNN) foi o que menos errou, em contraste com o método de aprendizado não-supervisionado (LDA) que teve o pior resultado nessa métrica.

Por fim, levando em consideração a Medida-F – por ser uma medida harmônica entre precisão e revocação, podemos concluir que o método CNN foi melhor no *Córpus* Buscapé. E que o método SF foi melhor no *Córpus* SemEval, seguido de perto do método CNN.

4.5 Ameaças à Validade

A validade de um instrumento pode ser definida como o âmbito em que as diferenças em escores observados do instrumento refletem as verdadeiras diferenças entre objetos quanto à característica que está sendo analisada. Nesta seção, destacamos duas variáveis que podem ameaçar a validade dos experimentos realizados, a saber:

- **Escolha de um método sobre outros que utilizam a mesma abordagem.** A escolha dos métodos que foram implementados para cada umas das quatro abordagens foi feita de forma empírica. Escolhemos os métodos de acordo com os trabalhos encontrados na literatura levando em conta a popularidade e resultados alcançados.
- **Escolha dos padrões linguísticos para a abordagem baseada em relações sintáticas.** Aqui nos utilizamos padrões já definidos em outros trabalhos que podem não ter alcançado resultados ótimos, uma vez que estamos usando um *Córpus* diferente do que o autor original utilizava, apesar de estar no mesmo domínio.

5 Considerações e Trabalhos Futuros

Essa Dissertação apresentou uma série de experimentos realizados para avaliar e comparar os resultados de quatro abordagens utilizadas para extrair aspectos na Mineração de Opiniões Baseada em Aspectos. Além disso, também comparou-se o comportamento dos métodos implementados entre *Corpora* de língua portuguesa e inglesa.

A abordagem comumente utilizada pelo nosso grupo de pesquisa se mostrou pouco eficiente ocupando as últimas colocações, mesmo sendo aplicados os padrões linguísticos de Sousa, Rabelo e Moura (2015) no mesmo domínio do seu trabalho. O modelo LDA demonstrou-se bem impreciso ao classificar aspectos, contraindicando usá-lo para esse fim. A abordagem baseada em frequência surpreendeu por ser um algoritmo simples e alcançar resultados melhores que as abordagens baseadas em modelos de tópicos e relações sintáticas. Já a abordagem usando aprendizado supervisionado alcançou resultados bem expressivos com uma precisão de até 96,76% no *Córpus* Buscapé, mas como a base de dados é bem desbalanceada a revocação foi baixa, apenas 35,73% no mesmo *Córpus*.

Este trabalho também criou um *Córpus* com 26.062 comentários de opiniões de consumidores sobre produtos em língua portuguesa com anotação semi-automática dos aspectos. Recurso muito escasso especialmente no idioma português, disponível em github.com/joaopaulolbq/corpusbuscape.

Outra contribuição é uma discussão sobre o uso de Redes Neurais Convolucionais para tarefas de classificação de sentimentos em nível de sentenças para comentários de língua portuguesa, que foi publicada em artigo (VIEIRA; MOURA, 2017).

5.1 Desafios e Limitações

A principal dificuldade encontrada foi a falta de recursos para a língua portuguesa, sobretudo em *Córpus* de aspectos anotado. Essa falta de recursos ocasionou na necessidade de criação de um novo *Córpus*. Outra dificuldade desta pesquisa foi o desbalanceamento inerente das classes *aspectos* e *não-aspectos* em relatos sobre produtos.

O *Córpus* foi criado de forma semi-automática e pode conter várias limitações como aspectos não anotados, aspectos implícitos, opiniões comparativas, palavras com grafia incorreta e ruídos. Outra limitação é que os modelos CNN e PL utilizados são dependentes do domínio da aplicação de relatos de produtos e serviços.

5.2 Trabalhos Futuros

Como trabalhos futuros, destaca-se:

- Aplicar o modelo CNN na dimensão extração de aspectos da abordagem Top(X) e variações.
- Analisar os modelos discutidos em outros domínios, tais como: hotéis, filmes, e redes sociais online.
- Combinar as técnicas SF e PL no modelo CNN para tentar melhorar a revocação.

Referências

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*. [S.l.: s.n.], 1994. p. 487–499. Citado na página 37.

ALUÍSIO, S. M.; PELIZZONI, J. M.; MARCHI, A. R.; OLIVEIRA, L. de; MANENTI, R.; MARQUIAFÁVEL, V. An account of the challenge of tagging a reference corpus for brazilian portuguese. In: *Computational Processing of the Portuguese Language, 6th International Workshop, PROPOR 2003, Faro, Portugal, June 26-27, 2003. Proceedings*. [S.l.: s.n.], 2003. p. 110–117. Citado na página 37.

ARCHAK, N.; GHOSE, A.; IPEIROTIS, P. G. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In: *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Jose, California, USA, August 12-15, 2007*. [S.l.: s.n.], 2007. p. 56–65. Citado na página 3.

BARBOSA, J. L. N.; MOURA, R. S.; SANTOS, R. L. de S. Predicting portuguese steam review helpfulness using artificial neural networks. In: *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web, Webmedia 2016, Teresina, Piauí State, Brazil, November 8-11, 2016*. [S.l.: s.n.], 2016. p. 287–293. Citado na página 31.

BECKER, K.; TUMITAN, D. *Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios*. [S.l.]: Anais do 28 Simpósio Brasileiro de Banco de Dados, 2013. 27–52 p. Citado 5 vezes nas páginas 15, 2, 5, 10 e 26.

BICKART, B.; SCHINDLER, R. M. Internet forums as influential sources of consumer information. *Journal of Interactive Marketing*, v. 15, n. 3, p. 31 – 40, 2001. ISSN 1094-9968. Citado na página 1.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. [S.l.]: O'Reilly, 2009. Citado na página 7.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. *Journal of Machine Learning Research*, v. 3, p. 993–1022, 2003. Citado 4 vezes nas páginas 18, 24, 25 e 30.

BONCHI, F.; CASTILLO, C.; GIONIS, A.; JAIMES, A. Social network analysis and mining for business applications. *ACM TIST*, v. 2, n. 3, p. 22:1–22:37, 2011. Citado na página 1.

BRANAVAN, S. R. K.; CHEN, H.; EISENSTEIN, J.; BARZILAY, R. Learning document-level semantic properties from free-text annotations. In: *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*. [S.l.: s.n.], 2008. p. 263–271. Citado na página 30.

CHEN, Y.; XIE, J. Online consumer review:Word-of-mouth as a new element of marketing communication mix. *Management Science*, v. 54, n. 3, p. 477–491, 2008. Citado na página 3.

CILIBRASI, R.; VITÁNYI, P. M. B. The google similarity distance. *IEEE Trans. Knowl. Data Eng.*, v. 19, n. 3, p. 370–383, 2007. Citado na página 28.

CLEVERT, D.; UNTERTHINER, T.; HOCHREITER, S. Fast and accurate deep network learning by exponential linear units (elus). *CoRR*, abs/1511.07289, 2015. Citado na página 19.

COLLOBERT, R.; WESTON, J.; BOTTOU, L.; KARLEN, M.; KAVUKCUOGLU, K.; KUKSA, P. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, JMLR.org, v. 12, p. 2493–2537, nov. 2011. ISSN 1532-4435. Citado 4 vezes nas páginas 22, 24, 26 e 29.

DELLAROCAS, C.; ZHANG, X. M.; AWAD, N. F. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, v. 21, n. 4, p. 23 – 45, 2007. ISSN 1094-9968. Citado na página 3.

GRAVES, A.; MOHAMED, A.-r.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE. *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. [S.l.], 2013. p. 6645–6649. Citado na página 22.

HOBBS, J. R.; RILOFF, E. Information extraction. In: *Handbook of Natural Language Processing, Second Edition*. [S.l.: s.n.], 2010. p. 511–532. Citado na página 12.

HOFMANN, T. Probabilistic latent semantic indexing. In: *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA, USA*. [S.l.: s.n.], 1999. p. 50–57. Citado 2 vezes nas páginas 18 e 30.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2004. (KDD '04), p. 168–177. ISBN 1-58113-888-1. Citado 5 vezes nas páginas 13, 16, 17, 27 e 28.

JAKOB, N.; GUREVYCH, I. Extracting opinion targets in a single and cross-domain setting with conditional random fields. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP 2010, 9-11 October 2010, MIT Stata Center, Massachusetts, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. [S.l.: s.n.], 2010. p. 1035–1045. Citado na página 29.

JIN, W.; HO, H. H. A novel lexicalized hmm-based learning framework for web opinion mining. In: *Proceedings of the 26th Annual International Conference on Machine Learning*. New York, NY, USA: ACM, 2009. (ICML '09), p. 465–472. ISBN 978-1-60558-516-1. Citado na página 29.

JINDAL, N.; LIU, B. Identifying comparative sentences in text documents. In: *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006*. [S.l.: s.n.], 2006. p. 244–251. Citado na página 11.

JONES, K. S. Natural language processing: A historical review. In: ___. *Current Issues in Computational Linguistics: In Honour of Don Walker*. Dordrecht: Springer Netherlands, 1994. p. 3–16. Citado na página 3.

KALCHBRENNER, N.; GREFENSTETTE, E.; BLUNSOM, P. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*, 2014. Citado na página 22.

KIM, H. D.; PARK, D. H.; LU, Y.; ZHAI, C. Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the Association for Information Science and Technology*, Wiley Online Library, v. 49, n. 1, p. 1–10, 2012. Citado na página 30.

KIM, S.; HOVY, E. H. Determining the sentiment of opinions. In: *COLING 2004, 20th International Conference on Computational Linguistics, Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland*. [S.l.: s.n.], 2004. Citado na página 12.

KIM, Y. Convolutional neural networks for sentence classification. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. [S.l.: s.n.], 2014. p. 1746–1751. Citado na página 22.

KINGMA, D. P.; BA, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. Nenhuma citação no texto.

KLAMBAUER, G.; UNTERTHINER, T.; MAYR, A.; HOCHREITER, S. Self-normalizing neural networks. *CoRR*, abs/1706.02515, 2017. Citado na página 19.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.* [S.l.: s.n.], 2012. p. 1106–1114. Citado 2 vezes nas páginas 21 e 22.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. [S.l.: s.n.], 2001. p. 282–289. Citado na página 29.

LI, M.; HUANG, L.; TAN, C.; WEI, K. Helpfulness of online product reviews as seen by consumers: Source and content features. *Int. J. Electronic Commerce*, v. 17, n. 4, p. 101–136, 2013. Citado na página 2.

LIU, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. [S.l.]: Springer, 2007. (Data-Centric Systems and Applications). Citado na página 9.

LIU, B. Sentiment analysis and subjectivity. In: *Handbook of Natural Language Processing, Second Edition.* [S.l.: s.n.], 2010. p. 627–666. Citado 5 vezes nas páginas 3, 9, 10, 11 e 12.

LIU, B. *Sentiment Analysis and Opinion Mining*. [S.l.]: Morgan & Claypool Publishers, 2012. (Synthesis Lectures on Human Language Technologies). Citado 10 vezes nas páginas 3, 9, 11, 12, 14, 15, 16, 26, 29 e 30.

LIU, B. *Sentiment Analysis - Mining Opinions, Sentiments, and Emotions*. [S.l.]: Cambridge University Press, 2015. Citado 3 vezes nas páginas 4, 8 e 26.

LIU, Q.; GAO, Z.; LIU, B.; ZHANG, Y. A logic programming approach to aspect extraction in opinion mining. In: *2013 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2013, Atlanta, GA, USA, November 17-20, 2013*. [S.l.: s.n.], 2013. p. 276–283. Citado na página 28.

LONG, C.; ZHANG, J.; ZHU, X. A review selection approach for accurate feature rating estimation. In: *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*. [S.l.: s.n.], 2010. p. 766–774. Citado na página 28.

MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. Distributed representations of words and phrases and their compositionality. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2013. p. 3111–3119. Citado na página 22.

MILNE, D. N.; WITTEN, I. H. An open-source toolkit for mining wikipedia. *Artif. Intell.*, v. 194, p. 222–239, 2013. Citado na página 1.

MOGHADDAM, S.; ESTER, M. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In: *Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM 2010, Toronto, Ontario, Canada, October 26-30, 2010*. [S.l.: s.n.], 2010. p. 1825–1828. Citado na página 27.

MOONEY, R. J.; BUNESCU, R. C. Mining knowledge from text using information extraction. *SIGKDD Explorations*, v. 7, n. 1, p. 3–10, 2005. Citado na página 12.

MUKHERJEE, A.; LIU, B. Modeling review comments. In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*. [S.l.: s.n.], 2012. p. 320–329. Citado na página 30.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21-24, 2010, Haifa, Israel*. [S.l.: s.n.], 2010. p. 807–814. Citado na página 19.

ORENGO, V. M.; HUYCK, C. R. A stemming algorithmm for the portuguese language. In: *Eighth International Symposium on String Processing and Information Retrieval, SPIRE 2001, Laguna de San Rafael, Chile, November 13-15, 2001*. [S.l.: s.n.], 2001. p. 186–193. Citado na página 37.

PARK, D.; LEE, J.; HAN, I. The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *Int. J. Electronic Commerce*, v. 11, n. 4, p. 125–148, 2007. Citado na página 3.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. [S.l.: s.n.], 2014. p. 1532–1543. Citado na página 22.

PEPPARD, J.; WARD, J. *The Strategic Management of Information Systems: Building a Digital Strategy*. [S.l.: s.n.], 2016. ISBN 9781119215479. Citado na página 2.

PONTIKI, M.; GALANIS, D.; PAVLOPOULOS, J.; PAPAGEORGIOU, H.; ANDROUTSOPOULOS, I.; MANANDHAR, S. Semeval-2014 task 4: Aspect based sentiment analysis. In: *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*. [S.l.: s.n.], 2014. p. 27–35. Citado na página 33.

PORIA, S.; CAMBRIA, E.; GELBUKH, A. F. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowl.-Based Syst.*, v. 108, p. 42–49, 2016. Citado 2 vezes nas páginas 18 e 29.

PORTEOUS, I.; NEWMAN, D.; IHLER, A. T.; ASUNCION, A. U.; SMYTH, P.; WELLING, M. Fast collapsed gibbs sampling for latent dirichlet allocation. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*. [S.l.: s.n.], 2008. p. 569–577. Citado na página 25.

PORTER, M. F. An algorithm for suffix stripping. *Program*, v. 14, n. 3, p. 130–137, 1980. Citado na página 37.

QIU, G.; LIU, B.; BU, J.; CHEN, C. Expanding domain sentiment lexicon through double propagation. In: *IJCAI 2009, Proceedings of the 21st International Joint Conference on Artificial Intelligence, Pasadena, California, USA, July 11-17, 2009*. [S.l.: s.n.], 2009. p. 1199–1204. Citado na página 18.

QIU, G.; LIU, B.; BU, J.; CHEN, C. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, v. 37, n. 1, p. 9–27, 2011. Citado na página 28.

RABINER, L. R. Readings in speech recognition. In: WAIBEL, A.; LEE, K.-F. (Ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990. cap. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, p. 267–296. ISBN 1-55860-124-4. Citado na página 29.

SÁ, C. A. de; VIEIRA, J. P. A.; MOURA, R. S. Approach to define author reputation in web product reviews using artificial neural networks. In: *2017 XLIII Latin American Computer Conference, CLEI 2017, Córdoba, Argentina, September 4-8, 2017*. [S.l.: s.n.], 2017. p. 1–10. Citado na página 31.

SANTOS, R. L. de S.; SOUSA, R. F. de; RABELO, R. A. L.; MOURA, R. S. An experimental study based on fuzzy systems and artificial neural networks to estimate the importance of reviews about product and services. In: *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*. [S.l.: s.n.], 2016. p. 647–653. Citado na página 31.

SANTOS, R. L. de S.; VIEIRA, J. P.; BARBOSA, J. L. N.; SÁ, C. A. de; MOURA, E. G. de B.; MOURA, R. S.; SOUSA, R. F. de. Evaluating the importance of web comments through metrics extraction and opinion mining. In: *35th International Conference of the Chilean Computer Science Society, SCCC 2016, Valparaíso, Chile, October 10-14, 2016*. [S.l.: s.n.], 2016. p. 1–11. Citado na página 31.

SARAWAGI, S. Information extraction. *Foundations and Trends in Databases*, v. 1, n. 3, p. 261–377, 2008. Citado na página 12.

SCAFFIDI, C.; BIERHOFF, K.; CHANG, E.; FELKER, M.; NG, H.; JIN, C. Red opal: product-feature scoring from reviews. In: *Proceedings 8th ACM Conference on Electronic Commerce (EC-2007), San Diego, California, USA, June 11-15, 2007*. [S.l.: s.n.], 2007. p. 182–191. Citado na página 28.

SHEN, Y.; HE, X.; GAO, J.; DENG, L.; MESNIL, G. Learning semantic representations using convolutionalneural networks for web search.In: *Proceedings ofthe 23rd InternationalConference on World Wide Web*. [S.l.: s.n.], 2014. (WWW '14 Companion), p. 373–374. ISBN 978-1-4503-2745-9.Citado na página 22.

SOUSA, R. F. de; RABELO, R. A. L.; MOURA, R. S. A fuzzy system-based approach to estimate the importance of online customer reviews. In: *2015 IEEE International Conference on Fuzzy Systems, FUZZ-IEEE 2015, Istanbul, Turkey, August 2-5, 2015*. [S.l.: s.n.], 2015. p. 1–8. Citado 5 vezes nas páginas 28, 31, 38, 43 e 45.

THET, T. T.; NA, J.; KHOO, C. S. G. Aspect-based sentiment analysis of movie reviews on discussion boards. *J. Information Science*, v. 36, n. 6, p. 823–848, 2010Citado na página 12.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. *Data Min. Knowl. Discov.*, v. 24, n. 3, p. 478–514, 2012Citado na página 4.

TURNEY, P. D. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of the 40th AnnualMeeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. [S.l.: s.n.], 2002. p. 417–424. Citado 3 vezes nas páginas 28, 38 e 43.

VIEIRA, J. P. A.; MOURA, R. S. An analysis of convolutional neural networks for sentence classification. In: *2017 XLIII Latin American Computer Conference, CLEI 2017, Córdoba, Argentina, September 4-8, 2017*. [S.l.: s.n.], 2017. p. 1–5Citado na página 45.

WIEBE, J.; WILSON, T.; CARDIE, C. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, v. 39, n. 2-3, p. 165–210, 2005Citado na página 12.

WILSON, T.; WIEBE, J.; HWA, R. Just how mad are you? finding strong and weak opinion clauses.In: *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence, July 25-29, 2004, San Jose, California, USA*. [S.l.: s.n.], 2004. p. 761–769Citado na página 12.

YIH, W.-t.; HE, X.; MEEK, C. Semantic parsing for single-relation question answering. In: *ACL (2)*. [S.l.: s.n.], 2014. p. 643–648Citado na página 22.

YU, J.; ZHA, Z.; WANG, M.; CHUA, T. Aspect ranking: Identifying important product aspects from online consumer reviews. In: *The 49th AnnualMeeting of the Association for ComputationalLinguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*. [S.l.: s.n.], 2011. p. 1496–1505Citado 2 vezes nas páginas 18 e 29.

ZHANG, L.; LIU, B. Identifying noun product features that imply opinions. In: *The 49th AnnualMeeting of the Association for Computational Linguistics: Human Language*

Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers. [S.l.: s.n.], 2011. p. 575–580. Citado na página 10.

ZHANG, L.; LIU, B. Aspect and entity extraction for opinion mining. In: *Data mining and knowledge discovery for big data*. [S.l.]: Springer Berlin Heidelberg, 2014. p. 1–40. Citado na página 11.

ZHANG, L.; LIU, B.; LIM, S. H.; O'BRIEN-STRAIN, E. Extracting and ranking product features in opinion documents. In: *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*. [S.l.: s.n.], 2010. p. 1462–1470. Citado na página 18.

ZHANG, Y.; WALLACE, B. C. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *CoRR*, abs/1510.03820, 2015. Citado 2 vezes nas páginas 15 e 23.

ZHUANG, L.; JING, F.; ZHU, X.-Y. Movie review mining and summarization. In: *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2006. (CIKM '06), p. 43–50. Citado na página 17.

ZÚÑIGA, H. Gil de; JUNG, N.; VALENZUELA, S. Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, Oxford University Press Oxford, UK, v. 17, n. 3, p. 319–336, 2012. Citado na página 1.