UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

André Lucas Machado Lins

Abordagem Híbrida e Independente de Domínio para extração de aspectos na Análise de Sentimentos

Recife

2018

André Lucas Machado Lins

Abordagem Híbrida e Independente de Domínio para extração de aspectos na Análise de Sentimentos

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Estatística e Informática

Curso de Bacharelado em Ciência da Computação

Orientador: Rinaldo Lima

Recife

2018

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

FICHA DE APROVAÇÃO DO TRABALHO DE CONCLUSÃO DE CURSO

Trabalho defendido por André Lucas Machado Lins às 14 horas do dia 16 de agosto de 2018, no Auditório do CEAGRI-02 – Sala 07, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado **Abordagem Híbrida e Independente de Domínio para Extração de Aspectos na Análise de Sentimentos**, orientado por Rinaldo José de Lima e aprovado pela seguinte banca examinadora:

Rinaldo José de Lima
DC/UFRPE

Adenilton José da Silva
DC/UFRPE

A meu pai Antônio Lins e mãe Rosalva Lins que mesmo com todas as dificuldades financeiras não mediram esforços para que eu pudesse chegar aqui…

Agradecimentos

Agradeço a Deus por todas as bênçãos que derrama todos os dias na minha vida, por não ter me desamparado e ter me disciplinado em todos os meus constantes erros.

Agradeço à minha família, em primeiro lugar à meus pais Antônio Lins e Rosalva Lins por ter me dado muito mais do que preciso, por terem investido todo o dinheiro que não tinham para me dar uma boa educação, por terem ficado até tarde várias vezes para me ajudar a estudar, por terem me criado na fé de Deus, sem eles eu não seria ninguém. Em segundo lugar agradeço a meu irmão Paulo Lins que esteve comigo e me apoiou em toda a minha caminhada.

Agradeço à minha namorada e futura esposa Daysnan Nicolly, por ter nos últimos 2 anos da faculdade sido fonte de animação para meus momentos difíceis, por ter me mostrado sempre o lado otimista de tudo, por ter me incentivado e ter sido companheira e amiga em toda essa caminhada.

Agradeço a meus amigos de faculdade, especificamente a turma de Ciência da Computação de 2014.1, obrigado por estarem comigo e me apoiar nos últimos anos, por me ajudar nas disciplinas e por trazer o senso de humor para todos os momentos, obrigado por permanecerem ao meu lado até o fim desse ciclo. Guardarei cada um de vocês no coração.

Agradeço a meus amigos Tamires e Cícero, por terem me ajudado a revisar todo esse trabalho, vocês foram fundamentais para a melhoria deste trabalho, obrigado pela amizade de vocês.

Agradeço a todos os meus professores, pelo conhecimento passado e por todos os seus conselhos, em especial ao meu orientador Rinaldo Lima por ter me apresentado a essa área e por ter conduzido meu fluxo de aprendizagem neste trabalho.

“ Porque Deus amou o mundo de tal maneira que deu o seu Filho unigênito, para que todo aquele que nele crê não pereça, mas tenha a vida eterna.”

(João 3:16)

Resumo

As opiniões são centrais a quase todas as atividades humanas e são chaves influenciadoras do nosso comportamento. Nossas crenças e percepções da realidade, e as escolhas que fazemos, são em grau considerável, condicionadas a como os outros veem e avaliam o mundo. Tendo em vista esta afirmação a área da Análise de Sentimentos ou Mineração de Opinião vem crescendo constantemente, a possibilidade de entender os sentimentos e opiniões que pessoas expressam sobre determinados assuntos enchem os olhos de todos. A Análise de Sentimentos(AS) é o estudo computacional das opiniões, atitudes e emoções das pessoas em relação a uma entidade. A literatura sobre Análise de Sentimentos é bastante vasta, existindo inúmeras variações de como realizar essa tarefa. Uma dessas variações da AS que vem recebendo bastante atenção dos pesquisadores nos últimos anos é a Análise de Sentimentos baseada em Aspectos(ASBA). Nessa abordagem os sentimentos são identificados em relação a aspectos de sentenças, a fim de discernir os tópicos que são tratados em cada sentença ou documento. A ASBA é dividida em três grandes tarefas que são a extração, classificação e agregação do aspecto, sendo a extração do aspecto como a tarefa mais complexa. Existem muitas abordagens para resolver a tarefa da extração de aspecto para ASBA, porém muitas dessas são abordagens dependentes de um domínio, o que dificulta replicar estas abordagens para outros domínios que não possuam as mesmas características. Logo, este trabalho visa propor um método híbrido e independente de domínio para extração de aspectos para ASBA, que consiste em quatro grandes etapas. A primeira identifica todos os aspectos candidatos a partir de regras semânticas para cada sentença. Após isso é gerado um léxico de todas as sentenças contendo os aspectos e sentimentos mais relevantes. Então segue-se a poda dos aspectos candidatos utilizando regras semânticas através do léxico de aspectos e sentimentos criados e, por último, é feita a seleção dos aspectos restantes através de um limiar dinâmico. Essa proposta foi avaliada nas bases de dados do Semeval 2016, contendo opiniões sobre vários aspectos relacionados com restaurantes e laptops, utilizando as métricas de avaliação mais utilizadas na literatura. Os resultados experimentais obtidos sugerem que o método proposto é competitivo quando comparado a vários outros métodos dependentes e independentes de domínio do estado da arte.

Palavras-chave: Análise de Sentimentos. Análise de Sentimentos baseada em Aspectos. Extração de Aspectos. Não Supervisionado. Análise Semântica.

Abstract

Opinions are central in most of the human activities and are keys of influence to our behaviors. Our beliefs, perception of reality and our choices are in a considerable degree, influenced by how people see and evaluate the world. In view of this statement the Sentiment Analysis(SA) has been growing constantly, the possibility of understand people's feelings and opinions about certain subjects gets everyone excited. Sentiment Analysis is the computational study of people's opinions, attitudes and emotions about some entity. The literature about Sentiment Analysis is pretty wide, having too many ways of execute such tasks. A variation of SA called Aspect based Sentiment Analysis (ABSA) has been receiving researchers attention. In this approach feelings are identified in relation to sentence aspects, in order to discern those that are treated in each sentence or document. ABSA is divided in three major tasks which are the extraction, classification and aggregation of the aspect, having aspect extraction as the most complex task. There's several approaches to solve the aspect extraction task, although many of these approaches are domain dependent, making difficult to replicate these approaches to domains that does not have the same features. Therefore, this work aims to purpose a domain independent hybrid method to aspects extraction, that consists in four major steps. The first one identify all the possible aspects out of semantic rules for each sentence. After this step, will be generated a lexical of all the sentences having the aspects and most relevant feelings. In the follow step is made the pruning of possible aspects using semantic rules through the lexical of aspects and feelings made previously. Lastly, is made a selection among the remaining aspects by a dynamic threshold. This purpose was evaluated in the Semeval's dataset, containing opinions about several aspects related to restaurants and laptops, using the most adopted evaluation metrics in literature. The experimental results imply that the proposed method is competitive when it's compared to many other methods dependents and independents of state of art domain.

Keywords: Sentiment Analysis. Aspect based Sentiment Analysis. Aspect Extraction. Unsupervised. Semantic Analysis.

Lista de ilustrações

Lista de tabelas

Lista de Algoritmos

*

Lista de abreviaturas e siglas

AA-Rel	Relação entre aspectos
AS	Análise de Sentimentos
AS-Rel	Relação entre sentimento e aspecto
ASBA	Análise de Sentimentos baseada em Aspectos
MO	Mineração de Opinião
PLN	Processamento de Linguagem Natural
PMI	Informações mútuas pontuais
POS	Part-of-speech
SS-Rel	Relação entre sentimentos
TF	Frequência de Termos
TF*IDF	Frequência de termos inverso frequência de documentos

Sumário

1 Introdução

As opiniões são centrais a quase todas as atividades humanas e são chaves influenciadoras do nosso comportamento (LIU, 2012). Nossas crenças e percepções da realidade, e as escolhas que fazemos, são em grau considerável, condicionadas a como os outros veem e avaliam o mundo (LIU, 2012). Por esta razão, quando precisamos tomar uma decisão ou mesmo fazer uma compra de um certo produto ou contratar um serviço, procuramos primeiro saber as opiniões dos outros a respeito (BANCKEN; ALFARONE; DAVIS, 2014). Ainda nesse contexto, saber a opinião ou sentimento de um grupo social pode ser de enorme importância para uma empresa lançar um novo produto ou gerenciamento de produtos existentes (PANG; LEE, 2008). Logo, saber o que os outros opinam ou determinar o sentimento sobre um determinado assunto é uma tarefa tanto interessante quanto valiosa.

Devido a isso, tem-se demonstrado um interesse cada vez maior em detectar, de forma automática, sentimentos ou opiniões a partir de documentos, para saber se esses sentimentos ou opiniões expressam polaridades positivas ou negativas sobre uma entidade. Porém, a conectividade quase ilimitada e um desejo enorme de compartilhar informações faz com que o volume de conteúdo de mídia social gerado pelos usuários cresça rapidamente (SCHOUTEN; FRASINCAR, 2016). Logo, ler e descobrir o sentimento de grandes quantidades de documentos textuais é difícil e demorado para um ser humano, tanto que, com o passar do tempo, torna-se quase que impossível se realizar tal classificação manualmente à medida que o tamanho da informação cresce. Assim, técnicas computacionais vêm sendo criadas para solucionar tal problema, surgindo então a área de pesquisa chamada Análise de Sentimentos (AS) (MULLEN; COLLIER, 2004), como uma subárea da Mineração de Textos.

A AS classifica sentenças e textos determinando se o usuário se posta de forma positiva, negativa ou neutra em relação a um determinado assunto (FELDMAN, 2013)(MEDHAT; HASSAN; KORASHY, 2014). Porém sabemos que em sua grande maioria, sentenças mais complexas (longas) podem ter opiniões positivas e negativas sobre um determinado assunto, em muitos casos o sentimento geral de um texto não é o mesmo que o de fragmentos descontextualizados, ou ainda, revisões negativas podem conter muitas frases aparentemente positivas, mesmo mantendo um tom fortemente negativo (STEINBERGER; BRYCHCíN; KONKOL, 2014).

Uma abordagem para tal detalhamento considera os termos principais ou mais relevantes de um documento e os analisa a fim de se ter uma melhor noção das características de um produto que o consumidor se interesse. Essa abordagem se chama

Análise de Sentimentos baseada em Aspectos (ASBA) (SCHOUTEN; FRASINCAR, 2016). Essa subárea de AS se divide em três tarefas principais: (i) a identificação e extração de aspectos, (ii) a classificação de aspectos e (iii) a agregação. A primeira consiste na identificação e extração de aspectos que se caracteriza por extrair os aspectos, isto é, termos ou conjuntos de termos que definem opiniões. Já a classificação de aspectos se caracteriza por atribuir os aspectos encontrados para um conjunto de categorias pré-definido. Por último temos a agregação que tem por objetivo unir vários aspectos em tópicos, a fim de dar uma visão de uma opinião de forma agregada. O foco desta pesquisa é na primeira tarefa, a identificação e extração de aspectos, que é considerada a tarefa mais complexa da ASBA (RANA; CHEAH, 2016).

Boa parte das metodologias do estado da arte propõe abordagens dependentes de domínio, isso significa que essas abordagens só funcionam em um determinado escopo, por exemplo o de produtos eletrônicos ou de hotéis, fazendo com que essas abordagens não possam ser reproduzidas em qualquer contexto. As abordagens que tentam ter domínio independente de domínio ainda possuem resultados bastante inferiores às outras. Assim, o objetivo deste trabalho é propor, projetar e implementar um sistema de extração de aspectos híbrido e independente de domínio para ASBA baseado na integração de técnicas estatísticas, baseado em regras e heurística de expansão.

1.1 Problema de Pesquisa

Apesar dos recentes avanços, a tarefa de identificação e extração de aspectos ainda está longe de ser resolvida devido ao fato das avaliações dos resultados não terem altas precisões para domínios independentes (FENG et al., 2014). Uma possível razão para isso é que os algoritmos existentes ainda são incapazes de lidar com frases complexas, que exigem mais do que a identificação de palavras portadoras de sentimentos ou por meio de simples análises de adjetivos que implicam em opiniões. No contexto da internet, existe um número ilimitado de maneiras que as pessoas podem usar para expressar opiniões (SCHOUTEN; FRASINCAR, 2016).

Uma das grandes dificuldades da ASBA é abordagens independentes de domínio, isso ocasiona bons resultados que não podem ser reproduzidos em outras bases de dados. Além disso, uma grande dificuldade na extração de aspectos na análise de sentimento é a extração de *keyphrases*, tal abordagem foi pouco utilizada na análise de sentimentos porém é ressaltada por alguns autores como uma linha de pesquisa importante (BAGHERI; SARAEE; JONG, 2014)(BAGHERI; SARAEE; JONG, 2013).

Na mineração de texto, a extração de *keyphrases* consiste em extrair um conjunto de frases relacionadas aos principais tópicos de um determinado documento(HASAN;

NG, 2014), sendo de enorme importância para áreas diversas da mineração de texto devido a possibilidade de que em um grande conjunto de documentos seja possível extrair os *keyphrases* mais relevantes.

As abordagens não supervisionadas são o grande foco na extração de aspectos na ASBA, constando com mais da metade das abordagens já propostas e os melhores resultados(RANA; CHEAH, 2016). Tipicamente, a tarefa de extração de aspectos para essas abordagens é composta por três etapas. A primeira etapa é a extração dos aspectos candidatos, isto é, extrair de um determinado documento os possíveis sintagmas que podem ser um possível aspecto de uma entidade qualquer. Posteriormente é feita a extração dos sentimentos ou opiniões relacionadas aos este aspectos candidatos e por último é utilizada alguma métrica de poda ou um limiar para escolher quais aspectos candidatos são realmente aspectos. Basicamente todas as abordagens (RANA; CHEAH, 2016) não supervisionadas seguem essas três etapas se baseando em apenas um algoritmo, seus resultados são bons em domínios dependentes, porém as que trabalharam com domínio independente, não obtiveram bons resultados.

Já foram propostas várias abordagens para extração de aspectos, primeiramente com as abordagens estatísticas (HU; LIU, 2004b), em seguida as abordagens probabilísticas (POPESCU; ETZIONI, 2007), posteriormente as abordagens de bootstrapping (BAGHERI; SARAEE; JONG, 2013) e por último e mais recentes abordagens baseadas em regras (PORIA et al., 2014). Todas essas abordagens obtiveram uma melhoria nos resultados relacionados a ASBA e principalmente a extração de aspectos. Porém como já foi mencionado boa parte destas abordagens ainda são dependentes de domínio, por exemplo, todas as abordagens citadas anteriormente são dependentes de domínio, sendo esses domínios normalmente de produtos eletrônicos. Marrese-Taylor, Velásquez e Bravo-Marquez (2013) ao propor uma metodologia que trabalha em cima do domínio de turismo, um domínio não muito comum na extração de aspectos na ASBA, tem seus resultados muito abaixo das outras abordagens. Todos os fatos citados nesse parágrafo mostram a necessidade de se dar uma maior ênfase no estudo de técnicas que sejam independente de domínio.

As abordagens apresentadas acima e os problemas que as envolvem mostram que ainda existe espaço para melhorias na extração de aspectos para ASBA, seja na independência de domínio ou o desenvolvimento de abordagens híbridas, logo, é proposto neste trabalho o seguinte problema :

1. Como desenvolver uma abordagem independente de domínio para extração de aspectos na ASBA, que independente do domínio, obtenha resultados próximos as abordagens já existentes na literatura e que dependem dos seus domínios?

A partir das abordagens já citadas neste projeto, tal proposta se baseia na hipó-

tese de que, quando usadas isoladamente, as técnicas citadas acima, tem suas vantagens e limitações. Porém, através de uma cuidadosa integração delas, este trabalho visa mitigar as desvantagens de cada uma delas, buscando assim uma maior sinergia entre elas e, ao mesmo tempo, melhorando sua acurácia na extração dos aspectos na AS.

1.2 Objetivos

Nesta seção estão expostos os objetivos propostos para este trabalho.

1.2.1 Objetivo Geral

Propor e desenvolver uma abordagem híbrida e independente de domínio para identificação e extração de aspectos na análise de sentimentos, que integre as técnicas de mineração de texto que são abordagens estatísticas, abordagens baseadas em regras sintáticas e abordagens de expansão.

1.2.2 Objetivos Específicos

1. Investigar o estado da arte em Análise de Sentimentos e, em particular, Análise de Sentimentos baseada em Aspectos.
2. Propor e implementar uma abordagem não-supervisionada e independente de domínio para o problema de extração de aspectos em textos.
3. Integrar várias técnicas e métodos já propostos na literatura afim de avaliar sua eficácia quando comparados com outros métodos já propostos.
4. Avaliar o desempenho da abordagem proposta usando bases de dados de competições e métricas adotadas na literatura.
5. Discutir vantagens e desvantagens da abordagem proposta em comparação com outras abordagens relacionadas no estado da arte.

1.3 Contribuições

Na busca pelos objetivos propostos, a abordagem utilizada neste trabalho visa contribuir para a ASBA da seguinte maneira:

- O aprimoramento e criação de regras semânticas para identificação de aspectos em qualquer tipo de textos sem considerar as características particulares de seu domínio.

- A implementação de um método híbrido, não-supervisionado e independente de domínio para a extração de aspectos na análise de sentimentos.
- Avaliação experimental e discussão de resultados que procuram ampliar o entendimento de abordagens híbridas não-supervisionadas quando aplicadas ao problema de extração de aspectos na ASBA.

1.4 Estrutura do Trabalho

Os próximos capítulos estão organizados da seguinte maneira: o Capítulo 2 apresenta os fundamentos básicos que envolve a teoria da Análise de Sentimentos, com foco na Análise de Sentimentos baseada em Aspectos.

O Capítulo 3 apresenta uma revisão da literatura, apresentando trabalhos relacionados a este trabalho, tanto auxiliando na metodologia como validando a importância desse trabalho, além de um resumo da evolução dos trabalhos da ASBA

O Capítulo 4 traz todo o método híbrido e independente de domínio para extração de aspectos na Análise de Sentimentos, apresentando como foi realizado a abordagem deste trabalho e o motivo das decisões, além de mostrar as mudanças e toda a evolução da abordagem.

No Capítulo 5 temos os experimentos feitos com bases de dados e métricas difundidas na literatura, as avaliações dos resultados deste trabalho e comparações em relação a outras abordagens da literatura. Por último o Capítulo 6 traz as conclusões do projeto e expectativas de trabalhos futuros do mesmo.

2 Fundamentação Teórica

2.1 Processamento de Linguagem Natural

O processamento de linguagem natural (PLN) é um componente da mineração de texto que realiza um tipo especial de análise linguística que essencialmente ajuda a máquina a compreender o texto. A PLN usa uma variedade de metodologias para decifrar as ambiguidades na linguagem humana, incluindo as seguintes: sumarização automática, marcação de parte da fala, desambiguação, extração de entidades e extração de relações, bem como a desambiguação, compreensão e reconhecimento da linguagem natural (MANNING et al., 2014). Na Figura 1 tem o funcionamento básico do Processamento de Linguagem Natural, onde em primeiro lugar o ser humano propõe textos. Esses textos são interpretados por uma ferramenta de PLN e essa ferramenta repassa os dados do texto de forma a ser compreendida pelo computador.

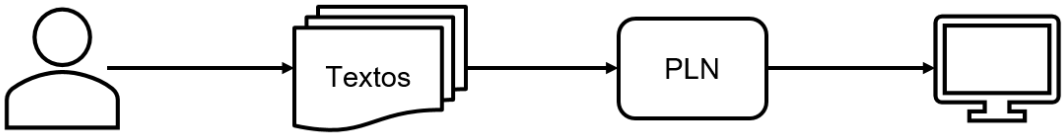

Figura 1 – Processamento de Linguagem Natural

2.1.1 Ferramentas de PLN

Ferramentas de PLN são utilizadas para adicionar informações não explícitas a um post. Estas informações são chamadas de anotações e requerem algum esforço computacional, dependendo da sua complexidade. Atualmente existem diversas ferramentas que oferecem um framework completo para o processamento de textos, dentre essas ferramentas temos como destaque o Stanford CoreNLP Toolkit (MANNING et al., 2014).

A ferramenta que este trabalho usa para processar as sentenças é o Standford CoreNlp que é uma estrutura de pipeline de anotações Java que fornece a maioria das etapas comuns do processamento de linguagem natural (PLN) (MANNING et al., 2014). O CoreNlp é sem dúvidas a ferramenta de processamento mais utilizada, trazendo uma gama de ferramentas como a análise léxica, tokenização e árvore de dependência, na Figura 2 é possível analisar toda a estrutura do CoreNlp.

Dentre as várias tarefas da Arquitetura do CoreNlp visto na Figura 2 podemos destacar a Análise Gramatical e a Análise Sintática. A Análise Gramatical tem

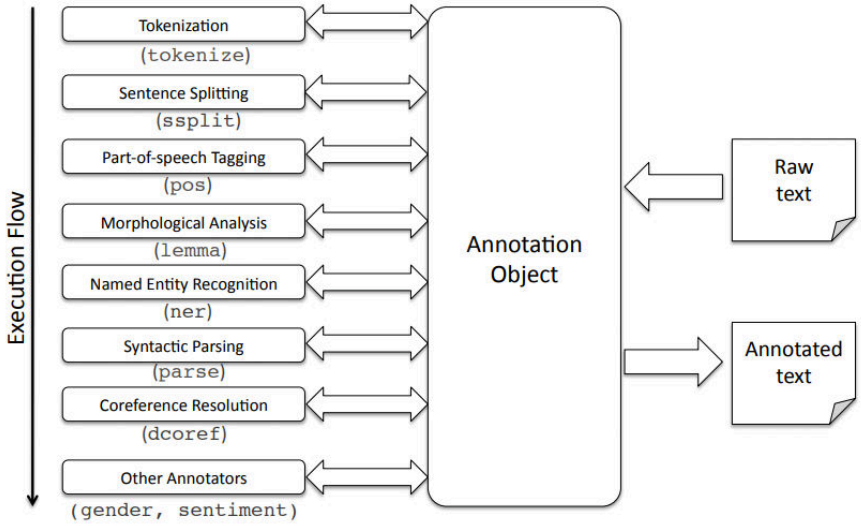

Figura 2 – Arquitetura do CoreNlp
(MANNING et al., 2014)

como objetivo a identificação das classes gramaticais, mais conhecidas como part-of-speech(POS). O CoreNLP identifica os POS através de simplificações, por exemplo, substantivo singular significa NN, na Tabela 1 temos todas essas regras gramaticais.

Nome	Descrição	Nome	Descrição
CC	Conjunções coordenativas	PRP$	Pronome possessivo
CD	Numeral cardinal	RB	Advérbio
DT	Delimitador	RBR	Advérbio comparativo
EX	”There”existencial	RBS	Advérbio superlativo
FW	Palavra estrangeira	RP	Palavras inflexivas
IN	Conjunções subordinativas	SYM	Símbolo
JJ	Adjetivo	TO	”To”como preposição
JJR	Adjetivo comparativo	UH	Interjeição
JJS	Adjetivo superlativo	VB	Verbo no infinitivo
LS	Marcador de item	VBD	Verbo no passado
MD	Verbo auxiliar	VBG	Verbo no gerúndio
NN	Substantivo singular	VBN	Verbo no particípio
NNP	Substantivo próprio	VBP	Verbo no presente
NNPS	Substantivo próprio plural	VBZ	Verbo na 3ª pessoa singular
NNS	Substantivo plural	WDT	”Wh”determinante
PDT	Pré-determinante	WP	”Wh”pronome
POS	Indicador possessivo	WP$	”Wh”pronome possessivo
PRP	Pronome pessoal	WRB	”wh”adverbial

Tabela 1 – Tabela de POS

A segunda tarefa é a Análise Sintática que tem como objetivo identificar a sequência de palavras e identificar as relações sintáticas entre essas palavras. As relações

sintáticas das palavras é o componente de maior importância no CoreNlp, através da árvore de dependências é possível identificar relações binárias entre as palavras. De acordo com Poria et al. (2014) essas relações são caracterizadas pelos seguintes recursos:

- O tipo da relação que especifica a natureza do link (sintático) entre os dois elementos na relação.
- O chefe da relação: este é o elemento que é o pivô da relação. As principais propriedades sintáticas e semânticas (por exemplo, concordância) são herdadas da cabeça.
- O dependente é o elemento que depende da cabeça e que geralmente herda algumas de suas características (por exemplo, número, gênero no caso de concordância).

Na Figura 3 temos um exemplo de como funciona a árvore de dependências do CoreNlp, sendo possível identificar as relações sintáticas, essas relações possuem siglas como “amod” que significa modificador adjetivo.

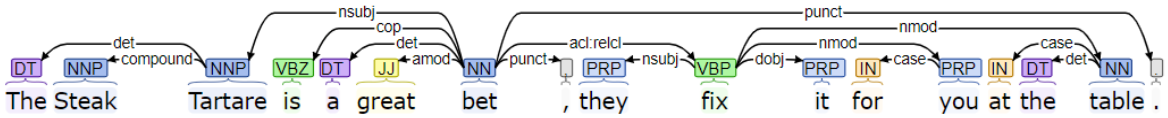

Figura 3 – Identificação das relações gramaticais do CoreNlp
<http://corenlp.run/>

As relações apresentadas na Figura 3, como “det”, “compound”, “nsubj” ou “cop”, são algumas das relações gramaticais possíveis. Entender essas relações é fundamental para maior precisão nos trabalhos da ASBA. Um exemplo disso é as relações de modificadores “amod” e “advmod”, que são respectivamente modificador ajetivo e adverbial, essas relações em sua grande maioria identificam um aspecto para a palavra que recebe esse modificador. Na Figura 3 existe a relação “amod” entre as palavras “*great*” e “*bet*”, onde a direção da seta identifica que o adjetivo “*great*” modifica o substantivo “*bet*”. Na tabela 2 tem uma lista das siglas e suas definições que identificam as relações gramaticais no CoreNLP.

2.2 Análise de Sentimentos

A Análise de Sentimentos (AS), também conhecida como Mineração de Opinião, é uma subárea da Mineração de Texto que consiste no estudo computacional das opiniões, atitudes e emoções das pessoas em relação a uma entidade (MEDHAT;

Nome	Descrição	Nome	Descrição
acomp	Complemento Adjetivo	npadvmod	Substantivo como Advérbio
advcl	Modificador Advérbio Clausal	nsubj	Sujeito Nominal
advmod	Modificador Adverbial	nsubjpass	Sujeito Nominal Passivo
agent	Agente	num	Modificador Numérico
amod	Modificador Adjetivo	number	Componente Numérico
appos	Modificador Apositivo	parataxis	Parataxe
aux	Auxiliar	pcomp	Complemento Preposicional
auxpass	Auxiliar Passivo	pobj	Objeto de uma Preposição
cc	Coordenação	poss	Modificador de Posse
ccop	Complemento Clausal	possessive	Modificador Possessivo
conj	Conjunção	preconj	Pré-Conjunção
cop	Cópula	predet	Pré-Determinador
csub	Sujeito Clausal	prep	Modificador Preposicional
csubjpass	Sujeito Passivo Clausal	prepc	Modificador Preposicional Clausal
dep	Dependência	prt	Modificador de Verbo Frasal
det	Determinador	pontuct	Pontuação
discourse	Elemento de Discurso	quantmod	Modificador de Quantidade
dobj	Objeto Direto	rcmod	Modificador Relativo Clausal
expl	Componente Expletivo	ref	Referência
goeswith	Acompanha	root	Raíz da Árvore
iobj	Objeto Indireto	tmod	Modificador Temporal
mark	Marcador	vmod	Modificador Redutivo Verbal
neg	Modificador de Negação	xcomp	Complemento Clausal Aberto
nn	Substantivo Composto	xsubj	Sujeito Controlador

Tabela 2 – Tabela de Relações Gramaticais

HASSAN; KORASHY, 2014), sendo essa entidade uma pessoa, objeto, tópico ou qualquer termo que se possa ter um opinião referente a ele. Além dos nomes citados acima a Análise de Sentimentos possui outros nomes como extração de opinião, mineração de sentimentos, análise de subjetividade, análise de afeto, análise de emoção, revisão de mineração.

Alguns autores tentam dividir todos os termos (extração de opinião, mineração de sentimentos, análise de subjetividade) citados acima (MEDHAT; HASSAN; KORASHY, 2014), relacionando por exemplo a Análise de Subjetividade a uma subtarefa da Mineração de Opinião (TSYTSARAU; PALPANAS, 2012). Logo, se foi criada uma definição que une de forma mais sucinta todos os pontos de vista dessa área de pesquisa, que é estudar os fenômenos de opinião, sentimento, avaliação, atitude e emoção (LIU, 2012). Neste trabalho entendemos que tanto a Mineração de Opinião como a Análise de Sentimentos possui o mesmo significado com modos de analisar diferente, portanto com frequência será mencionado opinião no lugar de sentimento.

Constando com cerca de 7000 trabalhos escritos (FELDMAN, 2013) em dife-

rentes contextos, a AS vem crescendo e tendo melhores resultados anualmente. Com todos esses trabalhos é bom ressaltar qual o objetivo final da AS, que é classificar os objetos de forma positiva, negativa ou neutra. Para isso foi criado várias formas e divisões de resolver o problema da análise de sentimentos e evoluir seus resultados, na Figura 4 temos uma visão de todas as áreas da Análise de Sentimentos.

A literatura consta com várias pesquisas sobre a Análise de Sentimentos que tentam explicar tudo o que envolve a AS. Mapeamentos sistemáticos como Feldman (2013), Schouten e Frasincar (2016), Cambria et al. (2013) e Montoyo, Martínez-Barco e Balahur (2012) trouxeram pesquisas fundamentais para AS. Liu (2012) traz um livro sobre o estado da arte da AS, e uma pesquisa bem avançada englobando todas as formas de realizar a análise de sentimentos ou mineração de opinião até o momento da sua publicação, este trabalho também apresenta as tendências de cada área e os algoritmos mais utilizados. Medhat, Hassan e Korashy (2014) traz uma pesquisa sobre a AS em nível de documento e sentença, apresentando a diferença entre as abordagens e o seu estado da arte. Rana e Cheah (2016) traz uma pesquisa sobre a AS em nível de aspecto, trazendo vários trabalhos sobre a ASBA e as tendências que a envolvem.

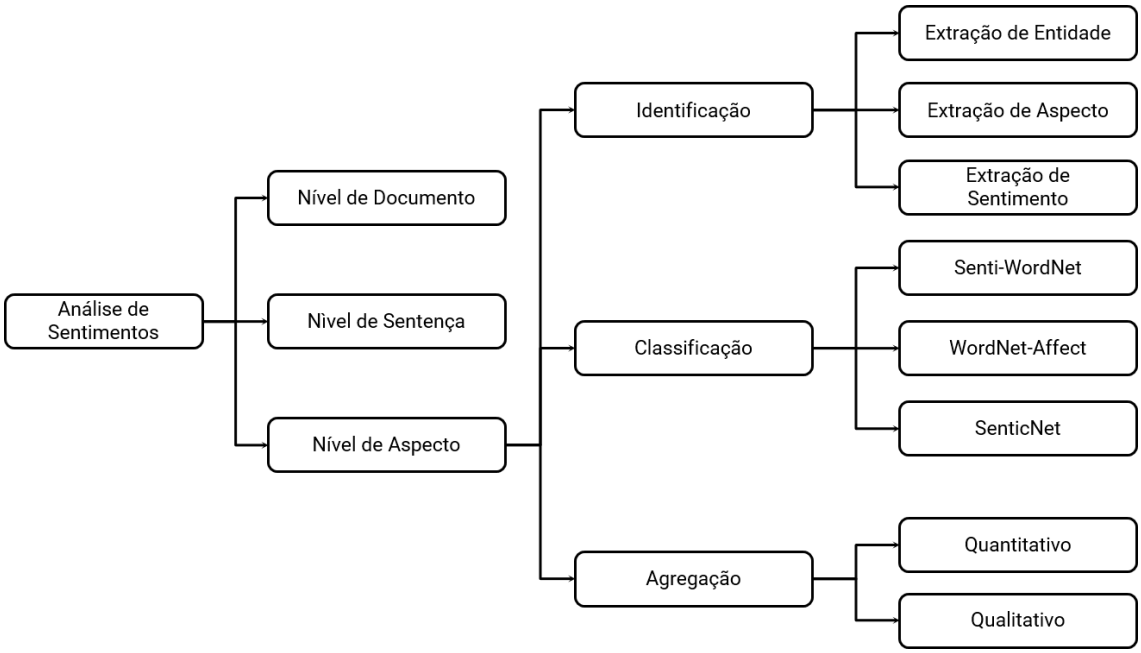

Figura 4 – Análise de Sentimentos
Adaptado de Rana e Cheah (2016)

A análise de sentimentos é um problema de PLN (LIU, 2012), por se relacionar a todos os aspectos da PLN. Questões como identificação da classe gramatical das palavras, relações semânticas e dependências gramaticais são tarefas da PLN que são fundamentais para a resolução das abordagens da AS. No entanto, alguns problemas importantes para AS ainda não são resolvidos na PLN, como a resolução de

referência, tratamento de negação e desambiguação de sentido de palavra, trazendo um maior grau de complexidade a AS. Vale ressaltar que a AS é um problema de PLN altamente restrito porque o sistema não precisa entender completamente a semântica de cada sentença ou documento, mas precisa apenas entender alguns aspectos dele, ou seja, sentimentos positivos ou negativos e suas entidades ou tópicos de destino. Nesse sentido, a AS oferece uma grande plataforma para os pesquisadores da PLN realizarem progressos tangíveis em todas as frentes da PLN, com o potencial de causar um enorme impacto prático (LIU, 2012).

2.2.1 Definições

Para um maior entendimento da AS, trazemos abaixo algumas definições de alguns termos que são frequentemente utilizados na AS e que serão bastante utilizados neste trabalho.

Sentimento O sentimento pode ser visto como um termo genérico para designar cada texto que expressa características positivas, negativas ou neutras. O termo "sentimento" é amplamente utilizado e pode se referir à subjetividade, emoção, avaliação e opinião.

Opinião De acordo com Liu e Zhang (2012), uma opinião (ou opinião regular) é um quíntuplo, que contém o nome de uma entidade, um aspecto da entidade, a orientação da opinião sobre o aspecto da entidade, o titular da opinião e o momento que acontece a opinião. A orientação da opinião pode ser positiva, negativa ou neutra, ou ser expressa com diferentes níveis de força ou intensidade.

Emoção Emoções são nossos sentimentos subjetivos e pensamentos, exemplo "Eu amo essa bebida" ou "Esse é o pior restaurante que já fui".

Entidade Uma entidade é um produto, serviço, tópico, questão, pessoa, organização ou evento. Ele é descrito com um par (T, W),onde T é uma hierarquia de partes, sub-partes e assim por diante, e W é um conjunto de atributos da entidade. Cada parte ou subparte também possui seu próprio conjunto de atributos (LIU, 2012).

Aspecto Aspecto é uma característica de uma entidade, por exemplo, "Qualidade da Câmera do Celular" é um aspecto da entidade "Celular".

2.2.2 Níveis da Análise de Sentimentos

Neste ponto traremos os níveis que a literatura utiliza para classificar as formas de fazer AS, os principais autores trazem três principais níveis, abaixo terá detalhes sobre esses três níveis.

2.2.2.1 Nível de Documento

A tarefa de SA em nível de documento é classificar se um documento de opinião inteiro expressa um sentimento positivo ou negativo (PANG; LEE; VAITHYANATHAN, 2002), considerando todo o documento uma unidade de informação básica (falando sobre um tópico). Por exemplo, dado uma crítica sobre uma música, a AS em nível de documento identifica se nesse documento a crítica foi positiva, negativa ou neutra em relação ao tópico do documento, que nesse caso foi a música. Como já foi ressaltado acima, esse nível de análise pressupõe que o documento tem um único tópico, não se aplicando a documentos que trazem opiniões e sentimentos sobre múltiplos tópicos ou unidades de informação.

2.2.2.2 Nível de Sentença

Neste nível a tarefa de SA classifica as sentenças de um documento, informando se elas expressam sentimento positivo, negativo ou neutro. O contexto de sentença está bastante ligado a classificação de subjetividade, neste caso é primeiro identificado se a sentença é subjetiva ou objetiva. As sentenças objetivas trazem informações factuais sobre um assunto, já as sentenças subjetivas informam visões e opiniões subjetivas. Comumente as expressões subjetivas são identificadas como as que expressam opiniões, porém alguns autores ressaltam frases onde mesmo sendo objetivas expressam uma opinião (LIU, 2012). Abaixo temos alguns exemplos de frases objetivas que trazem uma opinião referente.

"Nós compramos o carro no mês passado e o limpador de pára-brisa caiu."

No caso da frase acima mesmo sendo objetiva fica claro que existe uma crítica em relação ao para-brisa do carro, mostrando um descontentamento. Retirando a análise de subjetividade tarefa em nível de sentença não se difere da tarefa em nível de documento, pois uma sentença é só um documento menor (MEDHAT; HASSAN; KORASHY, 2014).

2.2.2.3 Nível de Aspecto

Como já dito na classificação em nível de documento, em caso de múltiplos tópicos não é possível ter uma classificação correta, isso também acontece em nível de sentença, para isso precisaremos trabalhar todas as palavras a fim de se discernir os tópicos que são tratados. Além disso tanto o nível de documento quanto o nível de sentença não possuem uma opinião sobre os aspectos e entidades de um tópico, não sendo possível saber exatamente as características que as pessoas gostaram ou não, logo para resolver os dois problemas foi criado a tarefa de SA que classifica os aspectos de um tópico. Análise em nível de aspecto, também chamado em nível

de recurso (HU; LIU, 2004a) ou entidade (LIU, 2012), tem como intuito classificar o sentimento em relação aos aspectos específicos da entidade.

Ao invés de olhar para construções de linguagem (documentos, parágrafos, frases ou cláusulas), o nível de aspecto analisa diretamente a própria opinião. Baseia-se na ideia de que uma opinião consiste em um sentimento (positivo ou negativo) e um alvo (de opinião) (LIU, 2012). Os detentores ou metas de opinião nos ajudam a entender melhor o problema da AS, abaixo temos duas frases que irão nos ajudar a entender melhor esse nível, uma com aparente tom negativo e outra com aparente tom positivo.

Embora o serviço não seja tão bom, eu ainda amo este restaurante

A qualidade deste telefone não é muito boa, porém a duração da bateria é longa

Na primeira frase percebemos um contexto positivo, porém ela não é totalmente positiva, isso também acontece na segunda frase onde mesmo demostrando um descontentamento com o telefone o usuário identifica um ponto positivo. O objetivo final desse nível é descobrir sentimentos em entidades e/ou aspectos a fim de classificar quantitativamente e qualitativamente essas entidades e aspectos.

A tarefa em nível de aspectos é a mais desafiadora dos níveis citados e tem vários problemas internos, essa será abordagem que este trabalho visa trabalhar, devido aos tantos problemas particulares e classificações que esse nível possui teremos um tópico mais a frente que irá tratar mais detalhadamente sobre esse nível de análise.

2.3 Análise de Sentimentos baseada em Aspectos

A Análise de Sentimentos baseada em Aspectos(ASBA) é referente ao terceiro nível da análise de sentimentos, nele é analisado os sentimentos referentes a cada aspecto de um documento ou sentença. Em geral três etapas de processamento podem ser distinguidas ao executar uma análise de sentimento em nível de aspecto: identificação, classificação e agregação (TSYTSARAU; PALPANAS, 2012; SCHOUTEN; FRASINCAR, 2016). Abaixo temos as responsabilidades destas 3 etapas.

Identificação Também conhecido como extração de aspectos, tem como objetivo extrair aspectos e sentimentos relacionados formando o par aspecto-sentimento, para que a partir desse par seja possível classificar o aspecto. Na frase *"A qualidade de voz deste telefone é incrível"*, notamos que existe o sentimento positivo "incrível" e que ele se relaciona a "qualidade de voz" que é o aspecto dessa sentença, formamos então o par de aspecto-sentimento. É importante notar na frase

anterior que seu tópico é o telefone, mas o aspecto não é o telefone e sim o termo no qual o sentimento se refere.

Classificação A classificação é uma tarefa posterior a Identificação que tem por objetivo a partir do par aspecto-sentimento classificar o aspecto a partir do sentimento em positivo, negativo ou em alguns casos neutro, No caso da frase citada acima *"A qualidade de voz deste telefone é incrível"*, já teríamos identificado o par "qualidade de voz" e "incrível", logo a classificação identifica que "incrível" é um sentimento positivo e identifica que o usuário trata "qualidade de voz" com sentimento positivo. Além dessa classificação em alguns trabalhos também é possível dar um peso para o sentimento, informando o grau de positividade e negatividade (RANA; CHEAH, 2016).

Agregação Um dos problemas que foi identificado anteriormente nas tarefas de AS em nível de documento e sentença era o fato que em múltiplos tópicos em um mesmo texto não havia a possibilidade de identificar corretamente o sentimento, a partir desse ponto é criado a agregação que a partir da identificação e classificação dos múltiplos aspectos agregamos eles em tópicos para uma identificação mais abrangente dos textos. Trabalhando ainda com a mesma frase dos outro tópicos mais acrescentando alguns detalhes temos a seguinte frase *"A qualidade de voz deste telefone é incrível e sua câmera é razoável"*, nesta frase percebemos outro par que é o aspecto "câmera" e o sentimento "razoável", assim iremos unir os dois pares para que seja possível ter uma visão mais concisa sobre o tópico que que eles se referem, que no caso é o telefone.

Além desses elementos centrais da análise de sentimentos em nível de aspecto, há outras preocupações: robustez, flexibilidade e velocidade (SCHOUTEN; FRASINCAR, 2016). Abaixo explicamos rapidamente essas três preocupações.

Robustez É muito comum o estilo de escrita informal, e quando tratamos do escopo de avaliação de itens, que é um contexto bastante comum na AS a escrita informal é bem mais frequente. Comumente as pessoas cometem erros de ortografia, utilizam gírias e siglas na escrita, além da utilização de *emoticons* na internet. Isso tudo torna a robustez uma preocupação muito importante para todos os trabalhos.

Flexibilidade É a capacidade de lidar com vários domínios. Os sentimentos e aspectos que são usados no domínio de telefones não são os mesmo usados no domínio de restaurantes. Alguns dos trabalhos que serão citados mais adiante possuem a problemática de só terem bons resultados para um domínio específico,

estes trabalhos obtêm ótimos resultados no domínio que se propõe a trabalhar, porém péssimos resultados em outros domínios.

Velocidade O intuito de toda AS é que a mesma possa ser utilizada para visualização de seus resultados para outros usuários, principalmente em interfaces web, logo a velocidade em que os dados são processados é muito importante.

Este trabalho visa aprimorar a primeira etapa da ASBA, que é a identificação ou extração de aspectos, sendo a etapa fundamental para que todas as outras duas etapas tenham resultados expressivos, sendo a tarefa mais desafiadora da ASBA (RANA; CHEAH, 2016). As preocupações que mais teremos neste trabalho e que serão melhoradas é a robustez e flexibilidade. No próximo tópico será enfatizado de forma mais detalhada a etapa de Extração de Aspectos.

2.3.1 Extração de Aspectos

Como já falado acima a extração de aspectos é a tarefa mais fundamental e desafiadora, dentro dela ainda existem algumas tarefas que estão resumidas na Figura 5. Os aspectos da Extração de Aspectos são separados pela primeira vez por HU; LIU, sendo divididos em Aspectos Implícitos e Aspectos Explícitos.

Aspecto Explícito São aqueles aspectos usados pelos usuários com palavras explícitas (RANA; CHEAH, 2016), por exemplo, na resenha: *"É leve o suficiente para levar com você para todos os lugares, mas poderoso o suficiente para obter fotos incríveis"*, o peso do aspecto foi expresso explicitamente. A extração de Aspectos Explícitos é amplamente estudada, contendo vários trabalhos relacionados a mesma.

Aspecto Implícito São aqueles aspectos que estão intrínsecos nas frases como por exemplo: *"É leve o suficiente para carregar o dia todo sem se preocupar"*, novamente o usuário fala sobre os aspectos do peso, porém desta vez nenhuma palavra explícita foi usada para expressar esses aspectos. A extração de aspectos implícitos é pouco estudada, possuindo poucos trabalhos relacionados a mesma, isso ocorre devido a alta complexidade de se extrair aspectos não informados em frases.

Dentro da extração de Aspecto Explícito podemos verificar na Figura 5 a divisão de Supervisionado, Não Supervisionado e Semi-Supervisionado, essa divisão leva em conta a necessidade de entradas de dados nos algoritmos, abaixo possui uma explicação sobre essas 3 maneiras de extrair aspectos explícitos.

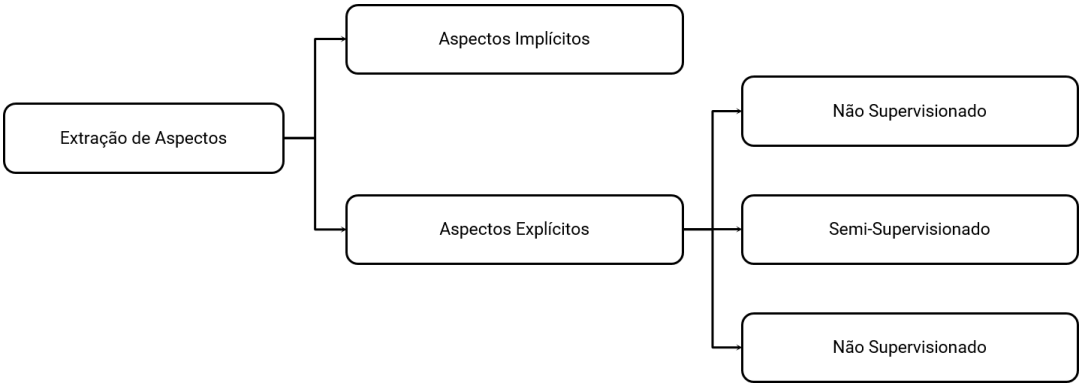

Figura 5 – Abordagens da Extração de Aspectos
Adaptado de Rana e Cheah (2016)

Algoritmo Não Supervisionado São técnicas que não dependem de entrada de usuário, isso significa que são livres de escopo. É comumente utilizado na revisão de produtos e avaliações online. Essa técnica corresponde a cerca de 46% dos trabalhos relacionados a extração de aspectos (RANA; CHEAH, 2016).

Algoritmo Semi-Supervisionado São técnicas que dependem parcialmente da entrada do usuário, precisando de sementes iniciais para o algoritmo. Assim como as técnicas de Algoritmos Não Supervisionado é também bastante utilizado nas revisões de produtos.

Algoritmo Supervisionado São técnicas que dependem totalmente da entrada do usuário, sendo normalmente restrita ao domínio que lhe propõe. Devido a boa parte dos datasets estarem na língua inglesa, os trabalhos se concentram mais nesta língua.

3 Revisão da Literatura

Como já foi ressaltado o objetivo deste projeto é a criação de uma metodologia não supervisionada para extração de aspectos, seguindo esse contexto trazemos uma lista de trabalhos relacionados a esse tema que serviram para maior entendimento da área, essa lista foi importada e adaptada do trabalho de (RANA; CHEAH, 2016), que traz um mapeamento sistemático do estado da arte da extração de aspectos para ASBA.

As abordagens não supervisionadas para extração de aspectos na ASBA foram divididas por (RANA; CHEAH, 2016) em quatro formas que são: Métodos Estatísticos, Métodos de Bootstrapping, Métodos baseados em Regras e Métodos Probabilísticos. Os trabalhos a seguir são divididos nessas quatro abordagens, e mostram a evolução dos trabalhos em cada abordagem e a importância destas abordagens para a abordagem proposta neste trabalho.

3.1 Métodos Estatísticos

Abordagens estatísticas foram as primeiras abordagens na mineração de texto e consequentemente as primeiras abordagens na análise de sentimentos, nesta seção teremos o detalhamento das abordagens de extração de aspecto que utilizam algoritmos estatísticos.

Hu e Liu (2004b) é o primeiro trabalho da literatura que visa a extração de aspectos, nele foi desenvolvido bases de dados para extração de aspectos que são utilizadas até os dias de hoje. O trabalho de HU; LIU tinha como objetivo extrair todos os aspectos mais frequentes e depois encontrar os sentimentos associadas a esses aspectos formando um par de aspecto sentimento, tal ideologia de um par de opinião é utilizada até os dias atuais em metodologias não supervisionadas. Estes aspectos frequentes eram substantivos e frases nominais que tinham maior quantidade de opinião de usuários, já esses sentimentos eram os adjetivos mais próximos. Em casos de palavras que não possuem substantivos ou frases nominais que são frequentes, é escolhido o substantivo mais próximo das palavras de opinião. O mesmo autor deste trabalho ainda propõe no mesmo ano o trabalho Hu e Liu (2004a) que tem como objetivo realizar as que seriam as futuras tarefas da AS, classificar os sentimentos e agregar os aspectos a fim de explicar o sentimento de uma sentença.

Raju, Pingali e Varma (2009) tem como objetivo extrair aspectos das revisões, para isso foi proposto uma abordagem em três etapas, pré-processamento, agrupa-

mento e extração de atributos. Primeiro eles extraem todos os substantivos e frases nominais no pré-processamento e, em seguida, classifica essas palavras através de *clusters*, a partir desses *clusters* é possível identificar os aspectos. O trabalho de RAJU; PINGALI; VARMA é independente de domínio e teve cerca de 52% de precisão média utilizando as bases de dados da <www.amazon.com>, identificando manualmente os aspectos das avaliações do website.

Ignorado por boa parte dos artigos anteriores a análise de websites fornece algumas informações adicionais, que são um conjunto de aspectos definidos e uma diretriz de classificação desses aspectos. Através dessas informações existentes, Moghaddam e Ester (2010) propõe o *Opinion Digger* que é uma abordagem não supervisionada para extrair aspectos das análises de clientes. Nessa abordagem, os autores usaram aspectos conhecidos de revisões do <http://www.epinions.com> para extrair aspectos explícitos das revisões de clientes. Essa abordagem é dependente do escopo e chegou a precisão de 77%.

Eirinaki, Pisal e Singh (2012) propõe o High Adjective Count(HAC) para extração de aspectos potenciais utilizando um sistema de pontuação. O EIRINAKI; PISAL; SINGH tem como objetivo extrair e classificar os aspectos, como o foco do nosso trabalho é a extração de aspectos, será focado somente nesse conhecimento. Este algoritmo encontra todos os substantivos no documento como aspectos potenciais atribuindo uma pontuação a cada um desses substantivos, e extrai os adjetivos como possíveis palavras de sentimento. Essa pontuação é atribuída através da quantidade de sentimentos que tem relação com esses aspectos e a força desses sentimentos através da frequência deles nas bases de dados. Os resultados desse trabalho variam de 40% à 70% nas bases de dados de revisão de produtos eletrônicos de HU; LIU.

Bafna e Toshniwal (2013) aprimorou o algoritmo de HU; LIU integrando uma abordagem baseada em probabilidade. É extraído todos os substantivos e frases nominais como aspecto, porém nem todos esses aspectos extraídos são de fato aspectos, para isso é utilizado a equação de poder probabilístico para remover todos esses substantivos que não representam aspectos. Após isso é extraído as palavras de sentimento, que são os adjetivos mais próximos dos aspectos extraídos.

Marrese-Taylor, Velásquez e Bravo-Marquez (2013) criou uma extensão para o domínio de turismo das técnicas já existentes de análise de sentimentos. Eles definem uma sentença como conjunto ordenado de palavras. A partir dessas palavras eles calcularam a distância entre duas palavras e usaram esse resultado da distância para a extração de aspectos. Essa abordagem foi mais tarde usada pelo próprio autor para criar uma ferramenta que analisou as avaliações do <www.tripadvisor.com>. Seus resultados na extração de aspectos tem média de 35%, sendo baixo em comparação com as outras abordagens.

3.2 Métodos de Bootstrapping

Outras abordagens de extração de aspectos são as abordagens baseadas em bootstrapping, nesta seção iremos ver abordagens, como é o seu funcionamento e a evolução dos trabalhos.

Bagheri, Saraee e Jong (2013) propõem o algoritmo de bootstrapping, que precisa de conjuntos iniciais de sementes de aspectos. Essa abordagem define um novo A-score matricial que utiliza informações de inter-relação entre palavras, e com isso uma lista de aspectos principais foi gerada. Esta informação de aspectos foi utilizada através do algoritmo de bootstrapping para gerar a lista final de aspectos utilizando o A-score de cada aspecto para mensurar a pontuação do valor, após isso é eliminado todos os aspectos redundantes. Essa abordagem ainda propõe dois métodos para realizar a poda e diminuir ainda mais as redundâncias, que são poda de suporte de subconjunto e poda de suporte de superconjunto. Essa abordagem é totalmente dependente de domínio, devido as sementes iniciais, e sua precisão é de 85%.

Li et al. (2015) propõem o SSPA, uma abordagem de bootstrapping baseada em padrões de dependência para extrair aspectos e opiniões. Para extração dos aspectos é utilizada regras semânticas através das relações entre substantivos e palavras de sentimentos. Após isso é utilizado o WordNet para agrupar aspectos semelhantes semanticamente, evitando redundâncias e filtrando os aspectos. Por último é calculada a reputação global de cada aspecto para descobrir a importância deles, ou seja, com que frequência os usuários estão comentando sobre estes aspectos.

3.3 Métodos Baseados em Regras(*Rule-Based*)

Abordagens baseadas em regras são uma das abordagens mais recentes e tem como objetivo utilizar regras semânticas, sintáticas ou modelos de tradução para identificar aspectos. Nesta seção traremos as abordagens não supervisionadas para extração de aspectos.

Liu et al. (2013) propõe o WTM, um modelo de tradução baseado em palavras para encontrar a associação entre aspectos e opiniões. Essa metodologia como várias das já mencionadas consideraram os substantivos e frases nominais como aspectos potenciais e adjetivos como seus sentimentos, porém diferente de boa parte das abordagens estatísticas, não foi considerado o adjetivo mais próximo como um sentimento, ao invés disso foi proposto um modelo baseado em gráficos para identificar as relações e então selecionar os aspectos e opiniões que têm a confiança. Os resultados desse trabalho tem precisão de cerca de 60% à 70%.

Htay e Lynn (2013) traz um método baseado em padrões para obter aspectos

e sentimentos. Esse método diferente dos vistos até agora utiliza advérbios e verbos além de frases nominais, substantivos e adjetivos para encontrar os padrões. Esse método extrai primeiro substantivos e frases nominais como aspectos, e adjetivos como sentimentos, formando pares com os mais próximos. Porém diferente dos outros métodos esse leva em conta os advérbios e verbos próximos de uma lista pré definida. Para avaliar essa abordagem os autores remove tuplas que contenham o menor número de verbos ou advérbios relacionados. Essa abordagem é dependente de escopo e chegou a ter precisão média de 80% na base de dados de HU; LIU.

Investigando a complexidade da sentença de revisão, Du, Chan e Zhou (2014) propõe o TrLM, um modelo de linguagem baseado em tradução, para identificar os aspectos do produto. Seu diferencial é que a abordagem indica que a qualidade das revisões também afeta as palavras de aspecto e sentimentos. A estrutura foi dividida em três subtarefas que são primeiro prever a importância de cada sentença de revisão utilizando o modelo de regressão do SVM. Na segunda parte, a informação reunida na primeira etapa foi incorporada ao modelo monolíngue de alinhamento de palavras, que extrai as relações de modificação entre os aspectos e as palavras de opinião. Na última etapa é realizada uma poda através de um limiar pré definido em cima da confiança dos aspectos definidos nos outros pontos.

Em grande maioria, os trabalhos anteriores, definiram as relações entre aspectos e sentimentos a partir da proximidade entre ambos, porém Poria et al. (2014) traz uma nova abordagem baseada em regras semânticas para extrair o aspecto explícito e implícito. Essas regras semânticas englobam várias classes gramaticais e se dividem em frases com sujeito substantivo e sem sujeito substantivo. Essa abordagem engloba tanto aspectos explícitos como aspectos implícitos, e utiliza o SenticNet (CAMBRIA; OLSHER; RAJAGOPAL, 2014) e o CoreNLP para identificação de sentimentos e análise de dependência das palavras. Essa abordagem também traz alguma regras mais aprimoradas de poda, sendo a abordagem totalmente baseada em regras mais completa até o momento. Essa abordagem não é livre de escopo e depende de uma entrada de aspectos explícitos e implícitos do escopo de produtos eletrônicos para treinar o algoritmo, sua precisão chega à 90%. O autor dessa abordagem ainda utiliza essas regras em Poria, Cambria e Gelbukh (2016) para extrair aspectos numa abordagem supervisionada.

3.4 Métodos Probabilísticos

Tendo em vista alguns problemas vindos das abordagens estatísticas como os casos onde substantivos pouco mencionados são normalmente aspectos (HU; LIU, 2004b), surgem as abordagens probabilísticas. Essas abordagens têm como objetivo

utilizar algoritmos e equações probabilísticas como o PMI e o LDA para extração de aspectos. Nesta seção iremos analisar as abordagens não supervisionadas probabilísticas para extração de aspectos.

Popescu e Etzioni (2007) baseado em HU; LIU, propõe uma modificação utilizando uma abordagem probabilística para uma maior precisão. Assim é removido os aspectos substantivos e frases nominais que não eram aspectos potenciais, removendo eles antes de ir para lista de aspectos extraídos. Para isso é utilizado o algoritmo probabilístico de informações mútuas pontuais(PMI), esse algoritmo avalia cada substantivo ou frase nominal que vá se tornar um aspecto, para tomar essa decisão é utilizado um limiar para identificar os possíveis aspectos que possuem um valor muito baixo, removendo eles da lista de aspectos extraídos.

Bagheri, Saraee e Jong (2014) propõe um modelo tópico probabilístico não supervisionado para extração de aspectos na análise de sentimentos, que visa extrair aspectos das revisões online levando em conta múltiplas palavras, esse modelo se chama ADM-LDA. Como o próprio nome informa, esse modelo é uma extensão do LDA, pois uni os parâmetros de diferentes documentos. Porém o mesmo não trata os documentos como saco de palavras, isso significa que o LDA não leva em consideração a posição de uma frase. Assim o autor propõe um algoritmo baseado em cadeias de Markov que leve em conta a posição das palavras, levando em conta que um aspecto de uma sentença é um provável aspecto de uma sentença futura. Avaliado em relação a outros algoritmos clássicos que utilizam a abordagem de modelos tópicos como o próprio LDA o mesmo traz resultados bastante positivos na extração de aspectos. Um de seus problemas é não encontrar simultaneamente o aspecto e seu sentimento, tendo isso como uma linha de pesquisa dita pelo autor. Além disso o autor focou em um algoritmo que seja independente de linguagem, assim não foi possível usar a semântica das palavras sendo mais difícil a extração de bons aspectos.

Quan e Ren (2016) traz uma união entre o PMI(abordagem probabilística) e o TF*IDF(abordagem estatística) Para encontrar a associação entre aspectos e aspectos de domínio. Aspectos de domínio são os aspectos no qual todo aspecto simples se refere, por exemplo no domínio de “câmera” o aspecto de domínio é “câmera”, assim essa abordagem indica que um aspecto tem que estar intimamente ligado ao aspecto do domínio. Logo, utilizando o mesmo domínio de “câmera”, “qualidade da imagem” é um aspecto, pois está intimamente ligado ao domínio de “câmera”. Após isso é utilizado o algoritmo PMI-TFIDF que mede a proximidade do aspecto extraído e seu sentimento. Por último é extraído os aspectos com maior valor de proximidade e que tenha alguma relação com o domínio. Essa abordagem é totalmente dependente de domínio e tem sua precisão média de 79%.

3.5 Outros métodos

Qiu et al. (2011) propõe um algoritmo baseado em dupla propagação e expansão do léxico. Além desses dois algoritmos é também usado o algoritmo CRF para essa abordagem. A dupla propagação consiste em realizar uma expansão do léxico de sentimentos e aspectos de forma conjunta, utilizando as relações de dependência direta e indireta entre aspectos e sentimentos. Essa abordagem é utilizada neste trabalho como forma de quantificar os aspectos através do léxico de aspectos e sentimentos, sendo assim será mais ressaltada em tópicos posteriores deste trabalho. Essa abordagem é supervisionada e totalmente dependente de domínio, tendo seus resultados com precisão média de 88%.

Lakkaraju, Socher e Manning (2014) propõe um novo conjunto de abordagens, neste conjunto é feita a extração de aspectos e seus sentimentos simultaneamente além de extrair múltiplos aspectos de sentenças. Para isso é utilizada aprendizagem de máquina com abordagem semântica. É provado que a adaptação do domínio pode facilitar a extração na Análise de Sentimentos. Sua técnica traz resultados melhores do que tecnologias de ponta. Tal técnica não leva em conta keyphrases tratando cada palavra como única.

Feng et al. (2014) propõe um método híbrido para extração de keywords na Análise de Sentimentos na linguagem chinesa. Nessa abordagem são utilizadas duas técnicas que são a análise semântica e análise de dependências sintáticas. A metodologia proposta é realizar um pré processamento, expandir o léxico de sentimentos em relação aos documentos, filtrar sentenças que baseadas no léxico possuam um sentimento, selecionar 4 características dessas sentenças sentimentais que são característica emocional, palavra-chave, dependência e posição e por último usar o SVM para classificar as sentenças como sentimentais ou não. Tal abordagem não leva em conta a extração de frases-chave, porém pode ser bastante útil para extração da mesma, tanto que o autor propõe como trabalho futuro a análise das estruturas de frase em sentenças. Outros pontos que autor ressalta como possíveis melhorias é a possibilidade de refinar o vetor de características e a possibilidade de usar um outro classificador além do SVM, para ver o impactos positivos que outros classificadores possam causar.

Brychcín, Konkol e Steinberger (2014) traz uma abordagem híbrida para extração e classificação de aspectos na análise de sentimentos utilizando inicialmente um sistema supervisionado baseado em aprendizagem de máquina e após isso usa um algoritmo não supervisionado(LDA) para descoberta de semântica latente. Tal abordagem foi usada para resolver a Task 4 da SemEval 2014[15], um workshop muito importante para avaliação de sistemas de mineração de textos em geral e, em particular, para a ASBA. Neste workshop o autor resolveu 4 subtarefas, porém a subtarefa 1 que é a extração de aspectos é a que importa para este projeto, com isso iremos

tratar apenas desse aspecto do trabalho de BRYCHCÍN; KONKOL; STEINBERGER. Para tal tarefa, o autor realizou mais um algoritmo probabilístico, o CRF(Conditional Random Fields), devido ao fato de tal técnica resolver a tarefa de reconhecimento de entidade nomeada. Os resultados encontrados foram superiores à média das técnicas já existentes, mostrando a eficiência da abordagem na Análise de Sentimentos(AS).

Liu et al. (2016) traz uma abordagem baseada em recomendações para melhorar duas formas de implementação de recomendação, a primeira baseada em semelhança semântica e a segunda em associação de aspectos. Essa abordagem é chamada de AER(Aspect Extraction based on Recomendation) e utiliza lifelong learning. A ideia por trás dessa abordagem é utilizar técnicas já conhecidas da extração de aspectos, adaptando as mesmas para os dados de recomendação e unindo-as. Essa abordagem traz uma característica importante a este projeto, que é o fato desta conseguir extrair multi palavras ou frases nominais que é o mesmo de *keyphrases*, devido ao fato deste utilizar o Standford Parser, assim essas frases são tratadas como substantivos nominais. Tal abordagem consegue resultados melhores em relação às técnicas separados em 3 de duas métricas, chegando a ter mais de 80% de precisão em algumas bases.

3.6 Resumo

Este capítulo apresentou abordagens para extração de aspectos no estado da arte da ASBA, apresentando algumas lacunas e direcionamentos a seguir na AS, como a independência de domínio, extração de *keyphrases* e as abordagens híbridas. Na Tabela 3 temos um resumo de todas essas abordagens, considerando as seguintes dimensões de estudo: algoritmo utilizado, modelo proposto e o domínio utilizado para avaliação.

Nos trabalhos acima podemos ver como se encontra o estado da arte da ASBA as dificuldades que existem e as lacunas. As abordagens não supervisionadas selecionadas correspondem a mais da metade das abordagens para extração de aspectos, as abordagens mais comuns são as abordagens estatísticas, bootstrapping, probabilísticas e baseadas em regras. Algumas dessas abordagens vem deixando de ser utilizadas individualmente, que é o caso das abordagens estatísticas, porém essas mesmas abordagens ainda sim participam de abordagens híbridas, sendo integradas com algumas das técnicas já mencionadas ou com algoritmos supervisionados. Os trabalhos mais recentes mostram que tem se aumentado o número de trabalhos com abordagem baseadas em regras (PORIA; CAMBRIA; GELBUKH, 2016) (LIU et al., 2016), se tornando uma tendência da extração de aspectos para ASBA. As propagações realizadas com abordagens de bootstrapping (LI et al., 2015) também tem se tornado mais

Referência	Modelo	Algoritmo	Domínio
Hu e Liu (2004b)	FBS	Frequency-based	Produtos
Popescu e Etzioni (2007)	OPINE	PMI	Produtos
Raju, Pingali e Varma (2009)	FB1*	Frequency-based	Produtos
Qiu et al. (2011)	Prop-dep	Double Propagation	Produtos
Moghaddam e Ester (2010)	Opinion Digger	Frequency-based	Produtos
Eirinaki, Pisal e Singh (2012)	HAC	Frequency-based	Restaurantes
Liu, Xu e Zhao (2012)	WTM	Word alignment + Graph-based	Vários
Bafna e Toshniwal (2013)	FPB	Frequency + Probability-based	Produtos
Marrese-Taylor, Velásquez e Bravo-Marquez (2013)	OZ	Frequency-based	Turismo
Bagheri, Saraee e Jong (2013)	BSTI*	Bootstrapping	Produtos
Htay e Lynn (2013)	PT-Based*	Pattern-based	Produtos
Bancken, Alfarone e Davis (2014)	ASPECTATOR	Syntatic dependency	Filmes e MP3
Du, Chan e Zhou (2014)	TrLM	Word alignment-based	Produtos
Poria et al. (2014)	Rule-based	Rule-based	Produtos
Lakkaraju, Socher e Manning (2014)	JMAS + RNTN	Deep Learning	Produtos
Feng et al. (2014)	SVM	SVM	Produtos
Brychcín, Konkol e Steinberger (2014)	UWB	LDA	Produtos
Liu et al. (2016)	AER	Lifelong Learning	Produtos
Quan e Ren (2016)	PMI-TFIDF	PMI	Produtos
Li et al. (2015)	SSPA	Bootstrapping	Produtos

Tabela 3 – Lista de Trabalhos com base na Extração de Aspectos para ASBA
Adaptado de Rana e Cheah (2016)

comuns, com muitas dessas abordagens já utilizando regras semânticas na propagação.

Uma lacuna que ficou perceptível em boa parte das abordagens é a independência de domínio, quase todos trabalhos citados na revisão apresenta abordagens dependentes de domínio. Isso se dá ao fato da dificuldade de se fazer a extração de aspectos para domínios independentes(RANA; CHEAH, 2016). As abordagens que possuem domínios que se diferem dos que já vem sendo trabalhados na literatura,

como o caso de Marrese-Taylor, Velásquez e Bravo-Marquez (2013) que trabalha com bases de dados de turismo, possuem resultados inferiores, quando comparados a outras abordagens. As abordagens que tentam ser independentes de domínio, como o caso de Raju, Pingali e Varma (2009), possuem resultados também inferiores as abordagens dependente de domínio.

Essa revisão literária ainda apresenta tendências da área que são as abordagens híbridas e a extração de aspectos *keyphrases*. A partir de que os autores começaram a utilizar as uniões de abordagens já existentes, a precisão dos seus resultados aumentam. Um exemplo é o caso de Quan e Ren (2016) que ao unir uma abordagem probabilística vindo do algoritmo PMI e uma abordagem estatística baseada em Hu e Liu (2004b) consegue resultados superiores a ambas as abordagens trabalhadas de forma separada. Isso acontece porque a abordagem estatística tem a dificuldade de analisar a importância das relações, a partir disso a abordagem probabilística entra para resolver essa problemática. Vemos em Feng et al. (2014) e (LIU et al., 2016) abordagens híbridas onde uma técnica supera as dificuldades das outras, combinando o algoritmo SVM, abordagem estatística ou propagação para trazer melhores resultados.

Devido às questões citadas nos parágrafos anteriores, que são a quantidade de abordagens que ainda dependem de domínio (RANA; CHEAH, 2016), e os resultados inferiores das abordagens independentes de domínio, como o caso de Raju, Pingali e Varma (2009). Este trabalho tem como objetivo apresentar uma abordagem de extração de aspectos independente de domínio para AS que através de uma integração das técnicas de regras semânticas, algoritmo estatístico e algoritmo de propagação, consiga melhorar os resultados para as abordagens independentes de domínio.

4 Proposta híbrida e Independente de Domínio para Extração de Aspectos

Neste capítulo é apresentado o método utilizado neste trabalho, fornecendo as informações necessárias para compreender a abordagem proposta. Também é apresentado a evolução da abordagem durante o estudo, trazendo todo o rigor metodológico que fez esse trabalho chegar até sua abordagem final.

A metodologia deste trabalho, trabalha em cima do domínio da língua inglesa, mas sendo possível a reprodução para outras línguas tendo em vista que as mesmas possuam uma ferramenta de processamento, base de dados e um conjunto de sentimentos que a linguagem se baseia. Este método é independente de domínio dentro de uma linguagem, sendo possível executar o algoritmo final em qualquer base de dados da língua inglesa.

4.1 Evolução do Método

O método proposto passou por várias etapas até chegar ao resultado proposto neste documento, essas etapas foram fundamentais para um amadurecimento da abordagem em relação ao domínio da extração de aspecto na ASBA.

Foi realizado um mapeamento da área de extração de aspectos para ASBA, a fim de descobrir o seu status atual e entender as tendências futuras. Tal mapeamento deixou claro algumas tendências como novas abordagens baseadas em grafos para realizar a extração de aspectos (LIU; XU; ZHAO, 2012). Outra abordagem que foi vista como uma tendência foi a utilização de regras semânticas para extração de aspectos como em Htay e Lynn (2013).

A partir dessas duas abordagens formamos a primeira abordagem híbrida, que tinha como objetivo integrar regras semânticas dentro de um grafo de dependências e utilizar o Topical PageRank para calcular os pesos dos vértices. Primeiro era identificado os aspectos candidatos que eram basicamente frases nominais e substantivos. Após isso foi utilizada a Árvore de Dependência do CoreNlp para criar um grafo das relações entre as palavras. Em terceiro lugar foi inserido as regras sintáticas de sujeito substantivo de Poria et al. (2014) para dar pesos as arestas. E por último foi utilizado o Topical PageRank para calcular o valor dos vértices, utilizando posteriormente um limiar para extração final dos aspectos. Essa técnica mesmo sendo bem embasada não proporcionou bons resultados e depois de várias tentativas de melhoria a precisão

não passava de 45%.

A primeira alteração para tentar melhorar a abordagem anterior foi extrair mais tipos de aspectos candidatos, para isso as regras de PORIA et al. se tornaram regras para extrair aspectos candidatos, ao invés de atribuir pesos para as relações e as relações seriam calculadas em relação ao peso sentimental que alguma das palavras expressam, esse peso vem do SenticNet (CAMBRIA; OLSHER; RAJAGOPAL, 2014). Porém tal mudança na abordagem não trouxe melhorias visíveis, fazendo com que a precisão só aumentasse 1%.

Tendo em vista as dificuldades de inserção de um algoritmo baseado em grafos optamos pela remoção do grafo e a inserção de uma propagação ou expansão de léxico. Para isso foi utilizado a abordagem (QIU et al., 2011), que propunha um algoritmo de dupla propagação para extração de aspectos. Essa abordagem utilizada como forma de avaliar os aspectos candidatos vindo do PORIA et al. fez com que a precisão subisse 10% nas primeiras tentativas. A partir disso, vendo o avanço em nível de resultado que a união dessas técnicas trouxe, essa abordagem se deu em torno de refinar e trazer mais técnicas para aprimorar os resultados.

4.2 Método Proposto

O método apresentado para Extração de Aspectos na ASBA consiste num modelo híbrido e independente de domínio que utiliza abordagens baseadas em regras semânticas, abordagens estatísticas e de expansão do léxico para se extrair os aspectos das sentenças. Uma visão geral dessa abordagem é apresentada na Figura 6 e abaixo dela temos um resumo dos pontos dessa abordagem.

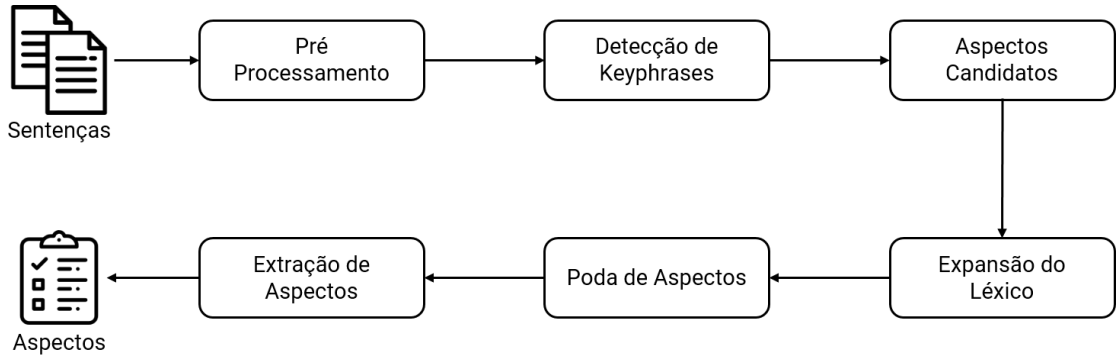

Figura 6 – Abordagem proposta para Extração de Aspectos
Elaboração Própria

1. **Pré-Processamento**: O primeiro passo dessa abordagem será a extração de aspectos, essa etapa se constitui em ler as sentenças e extrair os dados básicos e remover informações desnecessárias. Essa etapa tem como entrada todas as

sentenças das bases de dados, e tem como saída uma lista de palavras úteis para a abordagem. Essa etapa é constituída das técnicas de remoção de palavras de paradas (StopWords) e identificação da árvore semântica.

2. **Detecção de Keyphrases**: O segundo passo da abordagem é a detecção de *keyphrases* ou multi-word, que consiste na identificação de palavras que só fazem sentido quando unidas, transformando unigrams em bi-grams ou n-grams. Para isso foi utilizado algumas regras semânticas importadas e adaptadas de Bagheri, Saraee e Jong (2013).

3. **Aspectos Candidatos**: Esta terceira etapa tem como objetivo a detecção das palavras que devido ao seu contexto léxico possivelmente serão aspectos. Essa identificação é feita a partir de regras semânticas, sendo essas regras importadas dos trabalhos de (PORIA et al., 2014) e (PORIA; CAMBRIA; GELBUKH, 2016) que sofreram adaptações para serem livre de escopo, e outras regras foram definidas a partir da análise dos comportamentos.

4. **Expansão do Léxico**: Após a detecção dos aspectos candidatos, segue-se a expansão do léxico, nessa etapa será criado e expandido um léxico de sentimento e será refinado e expandido o léxico de aspectos criados. Essa etapa visa principalmente ser uma forma de avaliar o nível de importância de um aspecto candidato a partir do léxico de aspectos e sentimentos formado por todas as sentenças do documento.

5. **Poda de Aspectos**: A poda de aspectos candidatos refere-se ao refinamento e filtragem dos aspectos candidatos através de regras semânticas e análises estatísticas dos sentimentos encontrados.

6. **Extração de Aspectos**: A última etapa dessa abordagem é a extração dos aspectos a partir dos aspectos refinados até o momento. Essa etapa é fundamental para melhorar a precisão da abordagem, ela utiliza algoritmos estatísticos para classificar os aspectos das sentenças e após isso é criado um limiar para definir quais aspectos candidatos são de fato aspectos das sentenças.

4.2.1 Pré-Processamento

O Pré-Processamento tem por intuito de tratar o documento em linguagem natural não-estruturada, para um formato estruturado onde as palavras são identificadas individualmente (tokenização) bem como as análises sintáticas de funções gramaticais entre as palavras (árvore de dependência).

Para que seja possível a separação das palavras de uma sentença e análise léxica, é utilizado a ferramenta CoreNLP (MANNING et al., 2014), ferramenta já citada

na seção 2.1.1. Dado como entrada na ferramenta uma sentença, a ferramenta retorna a lista de palavras separadas, a marcação part-of-speech(POS) de cada palavra e as relações semânticas entre as palavras.

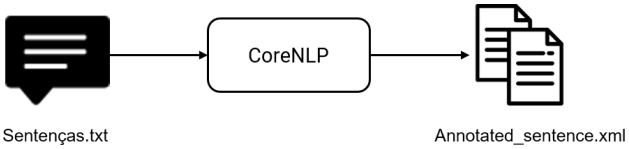

Figura 7 – Primeiro passo do Pré-Processamento

É detectado as unidades de palavra de cada sentença, o algoritmo identifica as palavras de parada, mais conhecidos como *stopwords*. Nossa lista de palavras de parada consta com 257 palavras da língua inglesa. Diferentemente de algumas metodologias, não ocorre a remoção, mas sim a identificação dessas palavras. Isso acontece devido ao fato de que algumas dessas palavras servem de ligação entre palavras importantes para ASBA.

Nessa etapa também é identificada as palavras que podem expressar sentimento, isso se dá através da base de dados Senticnet (CAMBRIA; OLSHER; RAJAGOPAL, 2014) que consta com 100 mil palavras de sentimentos, cada uma delas é acompanhada do seu grau. No Algoritmo consta um pseudocódigo da tarefa de Pré-Processamento.

Algoritmo 1: Pré-Processamento

```
Entrada: Sentença
Resultado: Palavras
início
    Palavras = Dividir Sentença;
    para todo Palavra em Palavras faça
        Identificar Relações Sintáticas;
        se Palavra consta em Stopword então
            Identificar Stopword;
        se Palavra consta em Senticnet então
            Identificar Sentimento;
```

4.2.2 Detecção de Keyphrases

Normalmente muitos dos aspectos possuem mais de uma palavra, por exemplo na frase *"Devido a péssima duração da bateria carrego meu celular todas as noites."*, onde observamos que o aspecto duração da bateria não é uma simples palavra. O aspecto não pode ser só a bateria pois a opinião do usuário é sobre um aspecto da

bateria que é a duração da bateria, outros exemplos são *"comida chinesa"*, *"sistema de carregamento do notebook"* ou *"Placa de Vídeo"*.

*I just bought the **digital camera** a few days ago. before I get used to it, here are my first feelings, the **picture quality** is so great, the **lens cover** is surely loose, the **zooming lever** is shaky , I hope it does not operate mechanically, otherwise you'll feel uneasy.*

Esta etapa do método tem como objetivo unir as palavras que só juntas possuem um significado, essas palavras são conhecidas como *keyphrases* (TURNEY, 2000) ou multi-word.

Consta-se com regras heurísticas para que seja possível identificar a necessidade de unir as palavras. Uma das formas mais simples é substantivos em sequência como por exemplo na frase *"Gosto de comida chinesa"*, "comida" e "chinesa" são dois substantivos em sequência logo os dois são uma única palavra. Existem também outras regras heurísticas que fazem essa detecção, Bagheri, Saraee e Jong (2013) traz uma lista de 4 regras para identificar essas palavras. A Tabela 4 explica o funcionamento dessas regras.

Descrição	**Regra**
Substantivo	Substantivos em sequência
Adjetivo e Substantivo	Adjetivos seguidos por substantivos
Determinante e Adjetivo	Determinantes seguidas por adjetivos
Substantivos e Verbos	Substantivos seguidos por verbos no gerúndio

Tabela 4 – Regras Heurísticas para detecção de keyphrases
Adaptado de Bagheri, Saraee e Jong (2013)

Essas regras auxiliam de forma significativa o encontro de *keyphrases* ou *multi-words*, porém elas podem identificar keyphrases que não fazem sentido, podendo ainda ter um substantivo interno ou uma keyphrase interna que pode identificar um aspecto mais relevante a frase. Dessas relações Bagheri, Saraee e Jong (2013) identifica a segunda regra da Tabela 4 como a regra onde é mais comum acontecer casos em que uma *keyphrase* não faz sentido, ele também propõe o limite *SubSet-Support* que tem por objetivo separar a parte adjetiva da substantiva e verificar se a segunda parte está significando mais que a keyphrase inteira. A seguir tem um exemplo dessa forma de avaliação.

$$Subset - Support(a) = \frac{Tf(a)+1}{Tf(a-1)+1}$$

Equação 1 - Subset-Support

Após toda a detecção de *keyhprases* é realizado o limite *Subset-Support* em todas as keyphrases encontradas, e caso esse limite seja menor que o valor 1 a *keyphrase* é repartida entre a parte adjetiva e parte substantiva, voltando a ser uma palavra individual, vale ressaltar que o *Subset-Support* é só para resolver os problemas da segunda regra da Tabela 4. No Algoritmo 2 consta um pseudocódigo da tarefa de Detecção de *Keyphrases*.

Algoritmo 2: Detecção de Keyphrases

```
Entrada: Palavras
Resultado: Palavras
início
    para todo Palavra em Palavras faça
        se POS(Palavra)=="Substantivo"e POS(Palavra+1) == "Substantivo"
          então
            Unir Palavras;
        se POS(Palavra)=="Adjetivo"e POS(Palavra+1) == "Substantivo" então
            Unir Palavras;
        se POS(Palavra)=="Artigo"e POS(Palavra+1) == "Adjetivo" então
            Unir Palavras;
        se POS(Palavra)=="Substantivo"e POS(Palavra+1) == "Verbo no
          Gerundio" então
            Unir Palavras;
        valor = Subset-Support(Palavra);
        se valor menor que 0 então
            Repartir Palavra;
```

4.2.3 Aspectos Candidatos

Esta terceira etapa da abordagem proposta tem por objetivo a detecção das palavras que devido ao seu contexto léxico e semântico possivelmente serão aspectos. Para realizar essa avaliação é usada regras sintáticas, essas regras são obtidas a partir do trabalho de Poria et al. (2014) que desenvolveu uma metodologia de extração de aspectos baseado em regras.

As regras desenvolvidas no trabalho de Poria et al. (2014) utilizam o SenticNet(CAMBRIA; OLSHER; RAJAGOPAL, 2014) para identificação de palavras de sentimentos e se baseia basicamente em duas regras gerais:

- Regras para as sentenças com verbo sujeito.
- Regras para as sentenças que não têm verbo sujeito.

Primeiro é identificado se uma palavra é um acionador de uma regra, isso significa que essa palavra é a primária na relação sintática. Após isso é calculado a contribuição, existindo várias maneiras que podem variar dependendo de como a relação de dependência e as propriedades das palavras correspondem às regras. A melhor forma é não considerar apenas a contribuição do acionador, mas em combinação com os outros elementos na relação de dependência(PORIA et al., 2014). Em primeiro lugar, o analisador do CoreNLP é usado para obter a estrutura de análise de dependência de cada sentença, a partir disso as regras de dependência criadas por Poria et al. (2014) e aprimorada no trabalho de Poria, Cambria e Gelbukh (2016) são empregadas nas árvores de análise para identificar aspectos candidatos.

O trabalho de Poria et al. (2014) leva em consideração também a extração de aspectos implícitos, se diferindo da abordagem deste trabalho que trabalha especificamente com a extração de aspectos explícitos. Com isso foi necessário a realização de algumas alterações nas regras desenvolvidas no trabalho de Poria et al. (2014), além disso algumas regras foram removidas e outras foram alteradas para as regras aprimoradas do trabalho de Poria, Cambria e Gelbukh (2016). Abaixo estão todas as regras utilizadas neste trabalho para identificar os aspectos, tais como seus exemplos.

4.2.3.1 Regras para Sujeito Substantivo

Para executar essas regras é necessário que uma palavra “t” esteja numa relação de sujeito substantivo com uma palavra “h”.

1. Se uma palavra “t” tem algum modificador adverbial ou adjetivo e o modificador existe no *SenticNet*, então “t” é identificado como um aspecto candidato.
2. Se a sentença não tiver verbo auxiliar, como por exemplo *should* ou *could*, então:
   - Se o verbo “t” é modificado por um adjetivo ou um advérbio, ou se está na relação de modificador de cláusula adverbial com outra palavra, então “h” e “t” são aspectos candidatos. No exemplo *battery* e *lasts* são aspectos candidatos.

     Exemplo: *The battery lasts little.*

   - Se “t” tem alguma relação de objeto direto com uma palavra “n” e o POS da palavra é um substantivo e não está no SenticNet, então “n” é identificado como um aspecto candidato. No exemplo *lens* é aspecto candidato.

     Exemplo: *I like the lens of this camera.*

- Se t tem alguma relação de objeto direto com uma palavra “n” e o POS da palavra “n” é substantivo e “n” existe no SenticNet, então a palavra “n” é extraída como um termo de aspecto. Na árvore de análise de dependência da sentença, se outra palavra “n1” estiver conectado a “n” usando qualquer relação de dependência e o POS de “n1” for substantivo, então “n1” será extraído como um aspecto. No exemplo *beauty* e *screen* são identificados como aspectos candidatos.

  Exemplo: *I like the beauty of the screen.*

3. Se a palavra “t” estiver em relação de cópula com um verbo copular e o POS de “h” for substantivo, então “h” será extraído como um aspecto explícito. No exemplo abaixo *camera* é identificada como um aspecto candidato.

   Exemplo: *The camera is nice.*

4.2.3.2 Regras para frases sem sujeito substantivo

Para as frases que não possuem relação de sujeito substantivo, os aspectos candidatos são identificados usando as seguintes regras.

1. Se uma palavra “h” estiver conectada a um substantivo “t” usando uma relação preposicional, ambos “h” e “t” serão extraídos como aspectos. No exemplo abaixo *camera* é identificada como um aspecto candidato.

   Exemplo: *Love the sleekness of the player.*

2. Se uma palavra “h” estiver em uma relação de objeto direto com uma palavra “t”, “t” será extraído como aspecto. No exemplo abaixo *price* é identificado como aspecto candidato.

   Exemplo: *Not to mention the price of the phone.*

4.2.3.3 Regras Adicionais

Por último também existem algumas regras que foram executadas a partir do resultado das regras anteriores.

1. Para cada aspecto candidato identificado nas outras regras, se um aspecto candidato “h” estiver em coordenação ou em relação conjunta com outra palavra

"t", então "t" também é identificado como um aspecto candidato. Na frase abaixo *amazing* já foi identificado como aspecto candidato, logo *use* também será um aspecto candidato.

Exemplo: *The camera is amazing and easy to use.*

2. Se "t" é identificado como um aspecto candidato e "t" possui um modificador composto de substantivo "h", então o aspecto "h-t" é identificado e "t" é removido da lista de aspectos candidatos. Na frase abaixo como o *chicken* e a *casserole* estão na relação modificador do substantivo composto, *chicken* é removido da lista de aspectos candidatos e *casserole* é adicionado.

   Exemplo: *We ordered the chicken casserole, but what we got were a few small pieces of chicken, all dark meat and on the bone*

4.2.4 Expansão do Léxico

A Expansão do Léxico ou Propagação do Léxico é uma etapa fundamental da metodologia deste trabalho, que tem como objetivo a criação do léxico de sentimento e a criação e expansão do léxico de aspectos. Esses léxicos irão servir para avaliar os aspectos candidatos extraídos na seção 4.2.3.

Essa expansão se baseia no trabalho de Qiu et al. (2011),que baseado em bootstrapping realiza uma propagação dupla nos documentos, é chamado de propagação dupla devido ao fato de que a propagação acontece em nível de sentimentos e aspectos, misturando as relações individuais e as relações mútuas. O método de Qiu et al. (2011) é considerado semi-supervisionado devido a utilização de bases de dados de sentimentos específico para os contextos de produtos tecnológicos, para que a metodologia deste trabalho não se torne semi supervisionado usaremos as palavras identificadas como sentimentos no capítulo 1, formando a partir dessas palavras o léxico de sentimentos inicial.

A base desta proposta é a identificação das relações, essas relações são entre as [1] sentimentos e outros sentimentos (SS-Rel), [2] aspectos e sentimentos (SA-Rel), [3] aspectos e outros aspectos (AA-Rel). Como já citado e para conveniência foram criadas as siglas SS-Rel, SA-Rel e AA-Rel que serviram para facilitar a pronúncia dessas relações, essas siglas se baseiam nas siglas de Qiu et al. (2011).

Para facilitar o entendimento das relações é criado dois tipos de dependências que são **Dependência Direta** e **Dependência Indireta**. Dependência direta indica que uma palavra depende da outra palavra sem quaisquer palavras adicionais em seu ca-

minho de dependência, a Figura 8 possui no grafo (a) e (b) exemplos desse tipo de dependência. Já a dependência indireta indica que uma palavra depende da outra palavra através de algumas palavras adicionais, a Figura 8 possui no grafo (c) e (d) exemplos desse tipo de dependência.

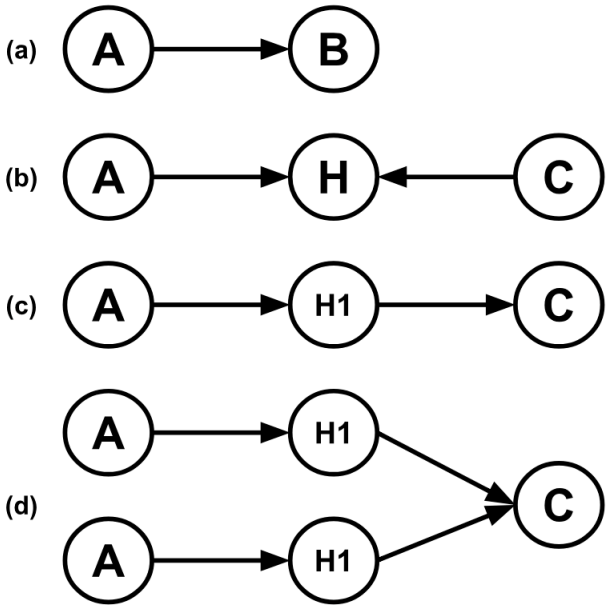

Figura 8 – Tipos de Dependência entre palavras

Através do estudo das relações, foi desenvolvido as regras que realizaram a expansão dos léxicos de sentimento e de aspectos. Essas tarefas são (1) extração de aspectos usando sentimentos (AS-Rel); (2) extração de aspectos usando os aspectos extraídos (AA-Rel); (3) extração de sentimentos usando aspectos (AS-Rel); (4) extração de sentimentos usando palavras de sentimentos extraídas(SS-Rel). Na metodologia deste trabalho, diferente do trabalho de (QIU et al., 2011), as tarefas são reorganizadas sendo primeiramente realizadas as regras (1) e (4) e posteriormente as regras (2) e (3), isso acontece devido ao fato de que as regras (1) e (4) dependem somente de um léxico de sentimentos inicial, tal léxico já foi desenvolvido na seção 1. Para cada uma dessas tarefas é criado regras que evidenciam as dependências de forma direta e indireta, na Tabela 5 está todas as regras sintáticas criadas a partir das tarefas citadas anteriormente.

Na Tabela 5 A é referente a Aspecto e S é referente a Sentimento, $\{A\}$ é referente ao léxico de Aspectos e $\{S\}$ é referente ao léxico de Sentimentos. Dep significa a dependência entre as palavras, onde $ADep$ é uma dependência com um Aspecto e $SDep$ é uma dependência com um Sentimento. Ainda nas regras temos $\{MR\}$ que significa as relações sintáticas mod, pnmod, subj, obj, obj2 e desc.

A metodologia deste trabalho executa neste ponto as 4 tarefas em todo o conjunto de documentos repetidas vezes até que não se surja nenhum novo aspecto ou

ID	Regra	Saída
R11	$S \rightarrow ADep \rightarrow A \Leftrightarrow S \in S, SDep \in MR, POS(A) \in NN$	Aspecto
R12	$S \rightarrow SDep \rightarrow H \leftarrow ADep \leftarrow A \Leftrightarrow S \in S, S/ADep \in \{MR\}, POS(A) \in NN$	Aspecto
R21	$S \rightarrow SDep \rightarrow A \Leftrightarrow A \in A, SDep \in \{MR\}, POS(S) \in \{JJ\}$	Sentimento
R22	$S \rightarrow SDep \rightarrow H \leftarrow ADep \Leftarrow A \Leftrightarrow A \in \{A\}, S/ADep \in MR, POS(S) \in \{JJ\}$	Sentimento
R31	$Ai(j) \rightarrow Ai(j)Dep \rightarrow Aj(i) \Leftrightarrow Aj(i) \in \{A\}, Ai(j)Dep \in \{CONJ\}, POS(Ai(j)) \in \{NN\}$	Aspecto
R32	$Ai \rightarrow AiDep \rightarrow H \leftarrow AjDep \leftarrow Aj \Leftrightarrow Ai \in \{A\}, AiDep == AjDep, POS(Aj) \in \{NN\}$	Aspecto
R41	$Si(j) \rightarrow Si(j)Dep \rightarrow Sj(i) \Leftrightarrow Sj(i) \in \{S\}, Si(j)Dep \in \{CONJ\}, POS(Si(j)) \in \{JJ\}$	Sentimento
R42	$Si \rightarrow SiDep \rightarrow H \leftarrow SjDep \leftarrow Sj \Leftrightarrow Si \in \{S\}, SiDep == SjDep, POS(Sj) \in \{JJ\}$	Sentimento

Tabela 5 – Regras Sintáticas para Expansão do Léxico
Adaptado de Qiu et al. (2011)

sentimento na nova iteração, gerando assim um léxico de aspectos e sentimentos que serão utilizados para avaliar os aspectos candidatos.

No Algoritmo 3 consta o pseudocódigo da R41, que é a regra na qual a expansão se inicia. Devido ao tamanho das regras só foi apresentada essa expansão.

Algoritmo 3: Expansão do Léxico na R41

```
Entrada: Sentenças
Resultado: Léxico de Aspectos, Léxico de Sentimentos
início
    para todo Sentença em Sentenças faça
        para todo Palavra em Palavras(Sentenças) faça
            se POS(Palavra)=="Adjetivo" então
                para todo Relação em Relações(Palavra) faça
                    se Tipo(Relação) == "CONJ"e Sentimento(Relação) ==
                     "verdadeiro" então
                        LexicoAspecto += Aspecto;
                    se Tipo(Relação) == "CONJ"e Sentimento(Relação) ==
                     "verdadeiro" então
                        LexicoAspecto += Aspecto;
```

4.2.5 Poda de Aspectos

É muito comum que durante uma abordagem baseada em regras nem todos os aspectos encontrados sejam de fato aspectos úteis, por isso que os aspectos que encontramos nas regras semânticas são chamados de Aspectos Candidatos. Devido a esses ruídos criados se torna necessário uma Poda desses Aspectos seguindo algumas heurísticas.

Logo a quinta etapa dessa abordagem é a poda de aspectos candidatos, uma etapa onde tem como foco refinar os aspectos através heurísticas e análises em relação ao léxico. Essas regras foram geradas a partir da análise dos resultados, verificando as melhorias que cada uma delas trazia para o resultado. Abaixo consta as heurísticas criadas.

- **Remoção de Stopwords**: *Stopwords* ou palavras de parada são palavras que podem ser consideradas irrelevantes para o conjunto de resultados relacionados a uma sentença. Essas palavras de parada são necessárias para a análise semântica porém em sua maioria não expressam aspectos interessantes, logo são ruídos que serão removidos dos aspectos candidatos.

- **Aspectos que não constam no léxico**: Um aspecto de uma sentença normalmente é um aspecto se o mesmo aparece outras vezes como aspecto em uma base de dados do mesmo contexto (QIU et al., 2011). Logo, a partir da criação de um léxico propagado de aspectos no conjunto de dados, desenvolvido na seção 3, será verificado se os Aspectos Candidatos são encontrados ao menos uma vez dentro do léxico propagado. Assim só teremos aspectos candidatos que são conhecidos no léxico da base de dados.

- **Aspectos após Comparações**: Quando as pessoas fazem comparações de aspectos, o aspecto que está sendo usado para fazer a comparação normalmente não é um aspecto que possui um sentimento para a sentença. Devido a isso é removido aspectos candidatos que estão posterior a palavras de comparação como *“compare to”* ou *“better than”*.

- **Aspectos sem relação com Sentimentos do Léxico**: Os sentimentos são normalmente parecidos dentro de uma base de dados, logo um aspecto de uma sentença normalmente é um aspecto se o mesmo se relaciona alguma vez com um sentimento que aparece como sentimento outras vezes na base de dados. Logo a partir da criação de um léxico propagado de sentimentos no conjunto de dados desenvolvido na seção 3, será verificado se os Aspectos Candidatos possuem ao menos uma relação com um sentimento comum no léxico de sentimentos.

Essas heurísticas são executadas na ordem que foram apresentadas e mostraram pequenas melhorias nos resultados. Na literatura já existem algumas outras formas de gerar uma poda em aspectos. No Algoritmo 4 consta um pseudocódigo da tarefa de Poda de Aspectos.

Algoritmo 4: Poda de Aspectos

```
Entrada: Aspectos Candidatos
Entrada: Lexico de Aspectos
Entrada: Lexico de Sentimentos
Resultado: Aspectos Podados
início
    para todo Aspecto em Aspectos Candidatos faça
        se Stopword(Aspecto) == verdadeiro então
            Remover Aspecto;
        se Aspecto não consta em Lexico de Aspectos então
            Remover Aspecto;
        se POS(Aspecto-1) == "Comparacao" então
            Remover Aspecto;
        se RelacoesSintaticas(Aspecto) não consta em Lexico de Sentimentos
         então
            Remover Aspecto;
```

4.2.6 Extração de Aspectos

A última etapa da abordagem proposta é a extração dos aspectos em cima dos aspectos refinados até o momento. Essa etapa é fundamental para se elevar a precisão da abordagem. Ela utiliza algoritmos estatísticos para classificar os aspectos das sentenças e após isso é definido um limiar para definir quais aspectos candidatos são de fato aspectos das sentenças.

Como já mencionado, essa etapa utiliza abordagem estatística para criar um peso em relação aos aspectos e após isso ser possível ordenar e passar um limiar para extração dos Aspectos. Esse peso é definido a partir da frequência do Aspecto Candidato dentro do léxico de aspectos e sentimentos desenvolvido na seção 3, e seu funcionamento é basicamente a frequência de termos(TF) do Aspecto Candidato dentro do léxico de aspectos somado a média dos TFs de todos as palavras relacionadas a esse Aspecto Candidato que constam no léxico de sentimentos da base de dados.

A equação a seguir expressa a forma de cálculo do peso de um aspecto candidato, onde “a” é um aspecto candidato, “s” é um sentimento e {S} é a quantidade de sentimentos relacionados a esse aspecto.

$$Peso(a) = Tfs(a) + \frac{\sum_{s=0}^{n} Tfs(s)}{n}$$

Equação 2 - Extração de Aspectos

A partir do cálculo do peso para cada aspecto candidato, é realizado uma ordenação de cada aspecto candidato em sua sentença referente, e após isso é passado um limiar escolhido manualmente para realizar a extração dos aspectos de cada sentença. No Algoritmo 5 temos um pseudocódigo ta tarefa de Extração de Aspectos.

Algoritmo 5: Extração de Aspectos

```
Entrada: Aspectos Candidatos
Entrada: Léxico de Aspectos
Entrada: Léxico de Sentimentos
Resultado: Aspectos
início
    para todo Aspecto em Aspectos Candidatos faça
        tf = tf(Aspecto);
        para todo Sentimento em Sentimentos(Aspecto) faça
            tfSent = tf(Sentimento);
        Peso(Aspecto) = tf + (tfSent / Count(Sentimentos(Aspecto)))
    Ordenar Aspectos Candidatos;
    para todo Aspecto em Aspectos Candidatos faça
        se Peso(Aspecto) menor que Limiar então
            Remover Aspecto;
```

4.3 Considerações Finais

Neste capítulo foi apresentado toda a abordagem deste trabalho, utilizando uma abordagem híbrida entre uma abordagem baseada em regras e uma abordagem de dupla propagação adaptando as mesmas para serem independente de domínio. Também foi utilizado algumas heurísticas para extração de *keyphrases* e uma equação para avaliação dessas *keyphrases*.

No próximo capítulo será mostrado os resultados obtidos a partir dessa metodologia, comparando as mesmas as metodologias usadas de forma separada e a metodologias independente de domínio.

5 Avaliação Experimental

Para que seja possível avaliar se os aspectos extraídos de um algoritmo são realmente úteis e correspondem ao que um ser humano entenderia como aspecto é necessário que este algoritmo seja avaliado. Neste capítulo será apresentado as bases de dados utilizadas nesta abordagem, as métricas de avaliação da extração de aspectos e em seguida será apresentado os experimentos realizados no método proposto, comparando o método proposto com outras abordagens da literatura.

5.1 Base de Dados

Um ponto fundamental de um trabalho na ASBA é a base de dados, essas bases de dados tem como intuito avaliar o desempenho de métodos de Análise de Sentimentos (SAIF et al., 2013). Neste trabalho foi utilizado a base de dados da Semeval, a base de dados mais utilizada nos trabalhos relacionados a extração de aspectos.

Semeval(Semantic Evaluation) é o Workshop internacional de Análise Semântica, que tem como objetivo através de suas bases de dados avaliar várias tarefas de Análise Semântica. Vem acontecendo anualmente desde 2012, porém existe desde 1998 com o nome Senseval. Para que se possa organizar melhor as tarefas que envolve as Análise Semântica, foi criado uma forma de divisão que consiste, *tasks* e *subtasks*, a segunda faixa tem como foco a Análise de Sentimentos e consta das seguintes tarefas:

- Tarefa 4: Análise de sentimentos no Twitter
- Tarefa 5: Análise de Sentimento Baseado em Aspectos
- Tarefa 6: Detectando Postura em Tweets
- Tarefa 7: Determinando a Intensidade do Sentimento das Frases em Inglês e Árabe

Como vimos acima uma dessas tarefas é a Análise de Sentimentos, atualmente sendo a tarefa 5 ela consta de uma série de subtarefas, sua segunda subtarefa é responsável pela extração de aspectos. A subtarefa de extração de aspectos é constituída de duas bases de dados que possuem um conjunto de avaliações de clientes, uma relacionada a comentários sobre Laptops e o segundo conjunto de dados contém dados relacionados a Restaurantes. Na Tabela 6 temos uma descrição dessas bases de dados.

Base de Dados	Domínio	Ano	Sentenças	Aspectos
Semeval	Laptops	2016	3308	5357
Semeval	Restaurantes	2016	2676	7745

Tabela 6 – Bases de Dados para Experimentos

Essas bases de dados vem no formato de XML onde cada item desta base consta a sentença, lista de aspectos relacionados a sentença e lista de categorias relacionadas a sentença. Na Figura 9 consta um exemplo de um item dessa base de dados, é possível notar todos os detalhes de um dado dessa base.

```
<sentence id="32897564#894393#2">
    <text>The bread is top notch as well.</text>
    <aspectTerms>
        <aspectTerm term="bread" from="4" to="9"/>
    </aspectTerms>
    <aspectCategories>
        <aspectCategory category="food"/>
    </aspectCategories>
</sentence>
```

Figura 9 – Exemplo de dado vindo da base de dados do Semeval

5.2 Métricas de Avaliação

Na extração de aspectos as principais métricas para avaliação dos algoritmos são precision, recall e f-measure (Hu e Liu (2004b), Popescu e Etzioni (2007), Bagheri, Saraee e Jong (2013), Quan e Ren (2016)). Tais métricas podem ser utilizadas para os 3 tipos de algoritmos para extração de aspectos que são supervisionado, semi-supervisionado e não supervisionado. Além disso todos os algoritmos citados no capítulo 3 utilizam pelo menos uma dessas métricas para avaliação..

Precision Denota a proporção de casos Positivos Previstas que são corretamente Positivos Reais (POWERS, 2007), isso significa que a precisão avalia o quanto o sistema acerta. Neste trabalho o precision avalia a quantidade de aspectos identificados corretamente, a Equação 3 é responsável pelo precision. Tp (true positive) representa o número de verdadeiros positivos, que na extração de aspectos são palavras que um algoritmo determina como aspecto. Enquanto fp (false positive) é o número de falsos positivos, que na extração de aspectos são palavras que um algoritmo determina como aspecto e que não são realmente aspectos.

$$\text{Precision} = \frac{tp}{tp + fp}$$

Equação 3 - Precision

Recall É a proporção de casos positivos que são corretamente previstos positivamente (POWERS, 2007), isso significa que o recall avalia a porcentagem de quantos itens relevantes foram de fato classificados como relevantes. Neste trabalho o recall avalia a quantidade de palavras que foram classificadas corretamente, a Equação 4 é responsável pelo recall. Tp (true positive) representa o número de verdadeiros positivos, que na extração de aspectos são palavras que um algoritmo determina como aspecto. Enquanto fn (false negative) é o número de falsos negativos, que na extração de aspectos são palavras que um algoritmo determina como não aspectos e que são realmente não aspectos.

$$\text{Recall} = \frac{tp}{tp + fn}$$

Equação 4 - Precision

F-measure É a média harmônica do precision e recall, que indica o quão próximo estão as métricas, tendo o seu melhor valor em 1 e o pior em 0. A Equação 5 é responsável pelo F-measure, relacionando o precision e recall.

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Equação 5 - F-measure

5.3 Avaliação da Abordagem Proposta

A proposta de extração de aspectos para ASBA explicada e desenvolvida no capítulo anterior é altamente personalizável e há vários componentes da abordagem que podem ser ajustados para obter um desempenho mais alto. Portanto, os experimentos desenvolvidos nesta seção tem o objetivo de usar a melhor configuração possível da abordagem proposta, utilizando as bases de dados da Semeval explicada nos capítulos anteriores.

Conforme explicado no Método deste trabalho, a abordagem proposta utiliza uma abordagem híbrida entre uma abordagem baseada em regras e uma abordagem de propagação do léxico, com o pressuposto que essas abordagens unidas trariam

melhores resultados que as estas mesmas abordagens separadas. Para a abordagem baseada em regras utilizamos o trabalho de Poria et al. (2014) como base e para a abordagem de propagação do léxico utilizamos o trabalho de Qiu et al. (2011) como base.

Porém um dos contextos da abordagem deste trabalho é que o mesmo funcione em diferentes domínios de base de dados, para isso as abordagens tiveram algumas modificações. Além disso as abordagens foram executadas em bases de dados diferentes dos que essa abordagem se propõe a utilizar, logo tais abordagens foram desenvolvidas separadamente e avaliadas.

Na Tabela 7 consta os resultados da abordagem híbrida proposta neste trabalho, em comparação com a abordagem baseada em regras de Poria et al. (2014) e a abordagem de propagação de léxico Qiu et al. (2011), alteradas para um contexto independente de domínio.

Semeval	**Regras**	**Propagação**	**Método Proposto**
Precision	45,5%	55,5%	61,2%
Recall	60,2%	62,3%	68,3%
F-Measure	51,8%	58,7%	64,6%

Tabela 7 – Resultado da Abordagem e comparação com Algoritmos não Híbridos

A abordagem proposta neste trabalho traz uma melhora na precisão em cerca de 14% em relação ao algoritmo baseado em regras e de 6% em relação ao algoritmo de propagação. A sua cobertura tem um aumento de 8% em relação ao algoritmo baseado em regras e de 5% em relação ao algoritmo de propagação.

Outro fator que este trabalho quis abordar foi a dependência de domínio das abordagens, em que boa parte das abordagens dependem de um domínio específico, como produtos ou hotéis, para executar seus algoritmos. Este trabalho propõe uma abordagem independente de domínio para a extração de aspectos, tendo em vista que a partir de que uma abordagem seja independente de domínio poderemos utilizar a mesma em qualquer contexto.

Existem poucos trabalhos que são independente de domínio em sua origem, para podermos comparar colocamos uma abordagem independente de domínio proposta no trabalho Raju, Pingali e Varma (2009) e outra abordagem que se propõe a utilizar bases de dados diferentes, proposta no trabalho Marrese-Taylor, Velásquez e Bravo-Marquez (2013). Na Tabela 9 temos os resultados propostos pelo método deste trabalho em comparação com as abordagens dos trabalhos de Raju, Pingali e Varma (2009) e Marrese-Taylor, Velásquez e Bravo-Marquez (2013).

Semeval	FB1	OZ	Método Proposto
Precision	51,5%	38%	61,2%
Recall	62,7%	48%	68,3%
F-Measure	56,6%	43%	64,6%

Tabela 8 – Resultado da Abordagem e comparação com Algoritmos Independentes de Domínio

A abordagem proposta neste trabalho traz uma melhora na precisão em cerca de 10% em relação ao FBI e de 20% em relação ao OZ. A sua cobertura tem um aumento de 5% em relação ao FBI e de 20% em relação ao OZ.

Por último é comparado os resultados desta abordagem independente de domínio com outras abordagens dependente de domínio. Esse tipo de comparação é importante para apresentar a relevância do método independente de domínio proposto, num contexto geral da literatura. Na tabela seguinte é apresentado essa comparação de resultado.

Semeval	Precision	Recall
Método Proposto	61,2%	68,3%
BST1	66,6%	84,5%
IEDR-Hotel	75,3%	47,3%
GP-Based	63,8%	99,8%
IERD	80,6%	80,5%

Tabela 9 – Resultado da Abordagem e comparação com Algoritmos Dependentes de Domínio

5.4 Discussão

Os resultados obtidos pela abordagem deste trabalho mostra que o mesmo tem resultados superiores em relação a outras abordagens independentes de domínio. Além disso tal abordagem se posta com resultados competitivos em relação a outras abordagens da literatura, tendo resultados comparativos com abordagens que são dependentes de domínio.Os resultados ainda mostram que essa abordagem traz melhores resultados que as abordagens simples que compuseram a nossa abordagem híbrida.

Podemos verificar através da Tabela 9 que a abordagem híbrida deste trabalho, em comparação com as abordagens separadas, possuem resultados melhores que as

abordagens utilizadas de forma separada. Isso confirma que ao criarmos abordagens utilizando união de abordagens já existentes e adaptando essas abordagens, a fim de uma suprir as deficiências da outra, será possível ter uma melhoria nos resultados já existentes na literatura.

Verificamos na Tabela 9 que o uso de uma abordagem híbrida trouxe melhores resultados para uma abordagem independente de domínio. Isso mostra que as abordagens híbridas podem ser a resposta para a grande dificuldade da extração de aspectos independente de domínio.

Observamos também na Tabela 9 a comparação dos resultados do método proposto, que é independente de domínio, em relação aos métodos dependentes de domínio. Constatamos que os resultados do método proposto se aproximam consideravelmente dos resultados de abordagens dependentes de domínio da literatura, tendo melhor cobertura que uma das abordagens. Isso encoraja a ideia que uma abordagem híbrida para o contexto de independência de domínio, apresenta resultados que se aproximam de outras abordagens que são dependentes de domínio. Essa afirmação pode trazer novos trabalhos híbridos e a realização de mudanças nos trabalhos da literatura para serem independentes de domínio.

6 Conclusão

Como apresentado neste trabalho, a extração de aspectos para a Análise de Sentimentos baseada em Aspectos que tem mais de 14 anos de pesquisa, é ainda um problema desafiador, com vários trabalhos publicados ao longo deste ano.

A revisão da literatura da ASBA mostrou que abordagens híbridas vem se tornando cada vez mais frequentes e que vem conseguindo apresentar melhores resultados quando visam explorar ao máximo as vantagens delas e ao mesmo tempo diminuir suas desvantagens quando aplicadas de forma isolada. Além disso boa parte dessas abordagens são dependentes de domínio, o que faz com que as mesmas só funcionem em bases de dados específicas, logo impossibilitando sua reprodução. Levando em conta tal discussão, o presente trabalho tem como objetivo a proposta e implementação de uma abordagem híbrida e independente de domínio para extração de aspectos para ASBA, apresentando a melhoria das abordagens híbridas em relação às abordagens simples e que mostre indícios que as abordagens híbridas podem trazer melhores resultados para abordagens independentes de domínio.

A abordagem proposta foi avaliada através das métricas de *precision*, *recall* e *f-measure*, e utilizou as bases de dados do semeval. Ela foi comparada com os algoritmos que a compõem de forma separada, e com outras abordagens que são independentes de domínio, a fim de compararmos as suas vantagens em relação a essas abordagens.

Os resultados obtidos encorajam novas abordagens a fazer uso de propostas híbridas para independência de domínio, tendo em vista os resultados positivos desta abordagem. Tais resultados mostraram indícios que as abordagens combinadas quando comparadas com as abordagens isoladas, trazem resultados satisfatórios, mostrando como a união de abordagens que se suprem trazem melhores resultados. Além disso essa abordagem foi comparada com algoritmos de extração de aspectos que são independentes de domínio, nessas comparações como na anterior os seus resultados foram superiores às abordagens independentes de domínio da literatura, encorajando que abordagens híbridas possam ser uma alternativa para solucionar o problema da dependência de domínio.

Em suma, a abordagem proposta pode ser vista como uma tentativa de uma abordagem híbrida, e independente de domínio para a extração de aspectos na ASBA, que possa trazer uma melhoria significativa nos resultados experimentais do estado da arte.

6.1 Trabalhos Futuros

A grande maioria das regras semânticas utilizadas para a identificação de aspectos candidatos na seção 4.2.3 foram inspiradas do trabalho de Poria et al. (2014) e adaptadas para o contexto independente de domínio. Porém, como o próprio autor ressalta, existe a possibilidade de melhorar tais regras semânticas através da definição ou refinamento de regras mais específicas para o problema em questão. Por conseguinte, uma primeira sugestão de melhoria do presente trabalho reside na adição de novas regras semânticas além daquelas que já se encontram nesse trabalho.

A etapa final do método proposto neste documento é relacionada a extração de aspectos. Nela formulou-se uma equação em função do TF dos termos presentes num corpus de documentos. Todavia, vários trabalhos apontam que tal métrica de ponderação da importância de termos não apresenta acurácia competitiva com outras tais como o PMI (POPESCU; ETZIONI, 2007). Por outro lado, o PMI necessita de uma grande quantidade de dados para ser mais efetivo (QUAN; REN, 2016). Dessa forma, trabalhos futuros poderão considerar a aplicação do método proposto em bases de dados de grande volume.

O foco do presente trabalho foi apenas a extração de aspectos explícitos. Por outro lado, seria interessante se investigar como a presente solução poderia ser adaptada para tratar também os aspectos implícitos. Muitas das técnicas existentes para extração de aspectos implícitos são dependentes de domínio, logo seria um grande desafio extrair aspectos implícitos e manter a abordagem independente de domínio.

Por último, a tarefa de extração de aspectos tem como sua tarefa posterior a classificação destes aspectos, identificando a polaridade dos mesmos. Dessa forma, uma extensão natural deste trabalho é prosseguir nas tarefas da ASBA, classificando os aspectos extraídos e em seguida determinando a polaridade em termos quantitativos dos aspectos.

Referências

BAFNA, K.; TOSHNIWAL, D. Feature based summarization of customers' reviews of online products. *Procedia Computer Science*, v. 22, p. 142 – 151, 2013. ISSN 1877-0509. 17th International Conference in Knowledge Based and Intelligent Information and Engineering Systems - KES2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050913008831>. Citado 2 vezes nas páginas 30 e 36.

BAGHERI, A.; SARAEE, M.; JONG, F. D. Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Know.-Based Syst.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 52, p. 201–213, nov. 2013. ISSN 0950-7051. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2013.08.011>. Citado 7 vezes nas páginas 14, 15, 31, 36, 40, 42 e 53.

BAGHERI, A.; SARAEE, M.; JONG, F. de. Adm-lda: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, v. 40, n. 5, p. 621–636, 2014. Disponível em: <https://doi.org/10.1177/0165551514538744>. Citado 2 vezes nas páginas 14 e 33.

BANCKEN, W.; ALFARONE, D.; DAVIS, J. Automatically detecting and rating product aspects from textual customer reviews. In: *Proceedings of the 1st International Conference on Interactions Between Data Mining and Natural Language Processing - Volume 1202*. Aachen, Germany, Germany: CEUR-WS.org, 2014. (DMNLP'14), p. 1–16. Disponível em: <http://dl.acm.org/citation.cfm?id=3053762.3053764>. Citado 2 vezes nas páginas 13 e 36.

BRYCHCÍN, T.; KONKOL, M.; STEINBERGER, J. Uwb: Machine learning approach to aspect-based sentiment analysis. In: *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*. Association for Computational Linguistics, 2014. p. 817–822. Disponível em: <http://www.aclweb.org/anthology/S14-2145>. Citado 3 vezes nas páginas 34, 35 e 36.

CAMBRIA, E.; OLSHER, D.; RAJAGOPAL, D. Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, 2014. (AAAI'14), p. 1515–1521. Disponível em: <http://dl.acm.org/citation.cfm?id=2892753.2892763>. Citado 4 vezes nas páginas 32, 39, 41 e 43.

CAMBRIA, E. et al. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, v. 28, n. 2, p. 15–21, March 2013. ISSN 1541-1672. Citado na página 22.

DU, J.; CHAN, W.; ZHOU, X. A product aspects identification method by using translation-based language model. In: *2014 22nd International Conference on Pattern Recognition*. [S.l.: s.n.], 2014. p. 2790–2795. ISSN 1051-4651. Citado 2 vezes nas páginas 32 e 36.

EIRINAKI, M.; PISAL, S.; SINGH, J. Feature-based opinion mining and ranking. *Journal of Computer and System Sciences*, v. 78, n. 4, p. 1175 – 1184, 2012. ISSN 0022-0000. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0022000011001139>. Citado 2 vezes nas páginas 30 e 36.

FELDMAN, R. Techniques and applications for sentiment analysis. *Commun. ACM*, ACM, New York, NY, USA, v. 56, n. 4, p. 82–89, abr. 2013. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/2436256.2436274>. Citado 3 vezes nas páginas 13, 21 e 22.

FENG, C. et al. A hybrid method of sentiment key sentence identification using lexical semantics and syntactic dependencies. In: HAN, W. et al. (Ed.). *Web Technologies and Applications*. Cham: Springer International Publishing, 2014. p. 11–22. ISBN 978-3-319-11119-3. Citado 4 vezes nas páginas 14, 34, 36 e 37.

HASAN, K.; NG, V. Automatic keyphrase extraction: A survey of the state of the art. v. 1, p. 1262–1273, 06 2014. Citado na página 15.

HTAY, S. S.; LYNN, K. T. Extracting product features and opinion words using pattern knowledge in customer reviews. v. 2013, p. 394758, 12 2013. Citado 3 vezes nas páginas 31, 36 e 38.

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. Disponível em: <http://doi.acm.org/10.1145/1014052.1014073>. Citado 2 vezes nas páginas 25 e 29.

HU, M.; LIU, B. Mining opinion features in customer reviews. In: *Proceedings of the 19th National Conference on Artifical Intelligence*. AAAI Press, 2004. (AAAI'04), p. 755–760. ISBN 0-262-51183-5. Disponível em: <http://dl.acm.org/citation.cfm?id=1597148.1597269>. Citado 9 vezes nas páginas 15, 27, 29, 30, 32, 33, 36, 37 e 53.

LAKKARAJU, H.; SOCHER, R.; MANNING, C. Aspect specific sentiment analysis using hierarchical deep learning. In: . [S.l.: s.n.], 2014. Citado 2 vezes nas páginas 34 e 36.

LI, Y. et al. A holistic model of mining product aspects and associated sentiments from online reviews. *Multimedia Tools and Applications*, v. 74, n. 23, p. 10177–10194, Dec 2015. ISSN 1573-7721. Disponível em: <https://doi.org/10.1007/s11042-014-2158-0>. Citado 3 vezes nas páginas 31, 35 e 36.

LIU, B. *Sentiment Analysis and Opinion Mining*. [S.l.]: Morgan & Claypool Publishers, 2012. ISBN 1608458849, 9781608458844. Citado 6 vezes nas páginas 13, 21, 22, 23, 24 e 25.

LIU, B.; ZHANG, L. A survey of opinion mining and sentiment analysis. In: ____. *Mining Text Data*. Boston, MA: Springer US, 2012. p. 415–463. ISBN 978-1-4614-3223-4. Disponível em: <https://doi.org/10.1007/978-1-4614-3223-4_13>. Citado na página 23.

LIU, K. et al. Opinion target extraction using partially-supervised word alignment model. In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*. AAAI Press, 2013. (IJCAI '13), p. 2134–2140. ISBN 978-1-57735-633-2. Disponível em: <http://dl.acm.org/citation.cfm?id=2540128.2540435>. Citado na página 31.

LIU, K.; XU, L.; ZHAO, J. Opinion target extraction using word-based translation model. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2012. (EMNLP-CoNLL '12), p. 1346–1356. Disponível em: <http://dl.acm.org/citation.cfm?id=2390948.2391101>. Citado 2 vezes nas páginas 36 e 38.

LIU, Q. et al. Improving opinion aspect extraction using semantic similarity and aspect associations. In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 2016. (AAAI'16), p. 2986–2992. Disponível em: <http://dl.acm.org/citation.cfm?id=3016100.3016320>. Citado 3 vezes nas páginas 35, 36 e 37.

MANNING, C. et al. The stanford corenlp natural language processing toolkit. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2014. p. 55–60. Disponível em: <http://www.aclweb.org/anthology/P14-5010>. Citado 3 vezes nas páginas 18, 19 e 40.

MARRESE-TAYLOR, E.; VELáSQUEZ, J. D.; BRAVO-MARQUEZ, F. Opinion zoom: A modular tool to explore tourism opinions on the web. In: *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. [S.l.: s.n.], 2013. v. 3, p. 261–264. Citado 5 vezes nas páginas 15, 30, 36, 37 e 55.

MEDHAT, W.; HASSAN, A.; KORASHY, H. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, v. 5, n. 4, p. 1093 – 1113, 2014. ISSN 2090-4479. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2090447914000550>. Citado 4 vezes nas páginas 13, 21, 22 e 24.

MOGHADDAM, S.; ESTER, M. Opinion digger: An unsupervised opinion miner from unstructured product reviews. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. New York, NY, USA: ACM, 2010. (CIKM '10), p. 1825–1828. ISBN 978-1-4503-0099-5. Disponível em: <http://doi.acm.org/10.1145/1871437.1871739>. Citado 2 vezes nas páginas 30 e 36.

MONTOYO, A.; MARTíNEZ-BARCO, P.; BALAHUR, A. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, v. 53, n. 4, p. 675 – 679, 2012. ISSN 0167-9236. 1) Computational Approaches to Subjectivity and Sentiment Analysis 2) Service Science in Information Systems Research : Special Issue on PACIS 2010. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0167923612001339>. Citado na página 22.

MULLEN, T.; COLLIER, N. Sentiment analysis using support vector machines with diverse information sources. In: . Barcelon, ES: [s.n.], 2004. Disponível em: <https://www.microsoft.com/en-us/research/publication/sentiment-analysis-using-support-vector-machines-with-diverse-information-sources/>. Citado na página 13.

PANG, B.; LEE, L. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1-2, p. 1–135, jan. 2008. ISSN 1554-0669. Disponível em: <http://dx.doi.org/10.1561/1500000011>. Citado na página 13.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: Sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2002. (EMNLP '02), p. 79–86. Disponível em: <https://doi.org/10.3115/1118693.1118704>. Citado na página 24.

POPESCU, A.-M.; ETZIONI, O. Extracting product features and opinions from reviews. In: ____. *Natural Language Processing and Text Mining*. London: Springer London, 2007. p. 9–28. ISBN 978-1-84628-754-1. Disponível em: <https://doi.org/10.1007/978-1-84628-754-1_2>. Citado 5 vezes nas páginas 15, 33, 36, 53 e 59.

PORIA, S.; CAMBRIA, E.; GELBUKH, A. Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems*, v. 108, p. 42 – 49, 2016. ISSN 0950-7051. New Avenues in Knowledge Bases for Natural Language Processing. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0950705116301721>. Citado 4 vezes nas páginas 32, 35, 40 e 44.

PORIA, S. et al. A rule-based approach to aspect extraction from product reviews. In: *SocialNLP@COLING*. [S.l.: s.n.], 2014. Citado 11 vezes nas páginas 15, 20, 32, 36, 38, 39, 40, 43, 44, 55 e 59.

POWERS, D. Evaluation: From precision, recall and fmeasure to roc, informedness, markedness and correlation. v. 2, p. 37–63, 01 2007. Citado 2 vezes nas páginas 53 e 54.

QIU, G. et al. Opinion word expansion and target extraction through double propagation. *Computational Linguistics*, v. 37, n. 1, 2011. Disponível em: <http://www.aclweb.org/anthology/J11-1002>. Citado 8 vezes nas páginas 34, 36, 39, 46, 47, 48, 49 e 55.

QUAN, C.; REN, F. Feature-level sentiment analysis by using comparative domain corpora. *Enterp. Inf. Syst.*, Taylor & Francis, Inc., Bristol, PA, USA, v. 10, n. 5, p. 505–522, jun. 2016. ISSN 1751-7575. Disponível em: <http://dx.doi.org/10.1080/17517575.2014.985613>. Citado 5 vezes nas páginas 33, 36, 37, 53 e 59.

RAJU, S.; PINGALI, P.; VARMA, V. An unsupervised approach to product attribute extraction. In: BOUGHANEM, M. et al. (Ed.). *Advances in Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 796–800. ISBN 978-3-642-00958-7. Citado 5 vezes nas páginas 29, 30, 36, 37 e 55.

RANA, T. A.; CHEAH, Y.-N. Aspect extraction in sentiment analysis: comparative analysis and survey. *Artificial Intelligence Review*, v. 46, n. 4, p. 459–483, Dec 2016. ISSN 1573-7462. Disponível em: <https://doi.org/10.1007/s10462-016-9472-z>. Citado 9 vezes nas páginas 14, 15, 22, 26, 27, 28, 29, 36 e 37.

SAIF, H. et al. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. In: *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*. [s.n.], 2013. Disponível em: <http://oro.open.ac.uk/40660/>. Citado na página 52.

SCHOUTEN, K.; FRASINCAR, F. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, v. 28, n. 3, p. 813–830, March 2016. ISSN 1041-4347. Citado 5 vezes nas páginas 13, 14, 22, 25 e 26.

STEINBERGER, J.; BRYCHCíN, T.; KONKOL, M. Aspect-level sentiment analysis in czech. 01 2014. Citado na página 13.

TSYTSARAU, M.; PALPANAS, T. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, v. 24, n. 3, p. 478–514, May 2012. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-011-0238-6>. Citado 2 vezes nas páginas 21 e 25.

TURNEY, P. D. Learning algorithms for keyphrase extraction. *Information Retrieval*, v. 2, n. 4, p. 303–336, May 2000. ISSN 1573-7659. Disponível em: <https://doi.org/10.1023/A:1009976227802>. Citado na página 42.