

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

ANDERSON UILIAN KAUER

**Análise de Sentimentos baseada em
Aspectos e Atribuição de Polaridade**

Dissertação apresentada como requisito parcial
para a obtenção do grau de Mestre em Ciência da
Computação

Orientador: Profa. Dra. Viviane P. Moreira

Porto Alegre
2016

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Kauer, Anderson Uilian

Análise de Sentimentos baseada em Aspectos e Atribuição de Polaridade / Anderson Uilian Kauer. – Porto Alegre: PPGC da UFRGS, 2016.

76 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR-RS, 2016. Orientador: Viviane P. Moreira.

1. Mineração de opinião. 2. Extração de aspectos. 3. Classificação de textos. 4. Análise de sentimentos. I. Moreira, Viviane P.. II. Título.

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Reitor: Prof. Carlos Alexandre Netto

Vice-Reitor: Prof. Rui Vicente Oppermann

Pró-Reitor de Pós-Graduação: Prof. Vladimir Pinheiro do Nascimento

Diretor do Instituto de Informática: Prof. Luis da Cunha Lamb

Coordenador do PPGC: Prof. Luigi Carro

Bibliotecária-chefe do Instituto de Informática: Beatriz Regina Bastos Haro

*“If I have seen farther than others,
it is because I stood on the shoulders of giants.”*

— SIR ISAAC NEWTON

AGRADECIMENTOS

Agradeço à minha família, em especial à minha mãe, agradeço eternamente pela minha educação e por seu exemplo de comprometimento e perseverança fez e continua fazendo com que seu filho nunca desista dos objetivos. Ao meu pai, da mesma forma, agradeço pelo exemplo de luta, humildade e honestidade. Aos meus irmãos pelo apoio e incentivo nestes dias intermináveis que foram necessários para que este trabalho se tornasse possível. À minha esposa, agradeço por seu apoio e suporte no dia-a-dia foram primordiais para que eu acreditasse que chegaria até aqui. Agradeço a meu grande amigo Marco Antônio de Assunção que foi extremamente importante para o desenvolvimento deste trabalho. Finalmente, agradeço a professora Dra. Viviane P. Moreira, minha orientadora, por acreditar em mim e propiciar a superação dos inúmeros momentos de dúvida, e aos demais professores do PPGC por contribuírem de alguma forma para o desenvolvimento do mesmo. Agradeço ao CNPq pelo incentivo à minha formação através da concessão de bolsa de estudos na modalidade de mestrado. Aos meus inúmeros colegas, especialmente Diego Costa Tumitan e Edimar Manica muito obrigado pelo companheirismo.

RESUMO

Com a crescente expansão da Web, cada vez mais usuários compartilham suas opiniões sobre experiências vividas. Essas opiniões estão, na maioria das vezes, representadas sob a forma de texto não estruturado. A Análise de Sentimentos (ou Mineração de Opinião) é a área dedicada ao estudo computacional das opiniões e sentimentos expressos em textos, tipicamente classificando-os de acordo com a sua polaridade (*i.e.*, como positivos ou negativos). Ao mesmo tempo em que sites de vendas e redes sociais tornam-se grandes fontes de opiniões, cresce a busca por ferramentas que, de forma automática, classifiquem as opiniões e identifiquem a qual aspecto da entidade avaliada elas se referem. Neste trabalho, propomos métodos direcionados a dois pontos fundamentais para o tratamento dessas opiniões: (*i*) análise de sentimentos baseada em aspectos e (*ii*) atribuição de polaridade. Para a análise de sentimentos baseada em aspectos, desenvolvemos um método que identifica expressões que mencionem aspectos e entidades em um texto, utilizando ferramentas de processamento de linguagem natural combinadas com algoritmos de aprendizagem de máquina. Para a atribuição de polaridade, desenvolvemos um método que utiliza 24 atributos extraídos a partir do ranking gerado por um motor de busca e para gerar modelos de aprendizagem de máquina. Além disso, o método não depende de recursos linguísticos e pode ser aplicado sobre dados com ruídos. Experimentos realizados sobre *datasets* reais demonstram que, em ambas as contribuições, conseguimos resultados próximos aos *baselines* mesmo com um número pequeno de atributos. Ainda, para a atribuição de polaridade, os resultados são comparáveis aos de métodos do estado da arte que utilizam técnicas mais complexas.

Palavras-chave: Mineração de opinião. extração de aspectos. classificação de textos. análise de sentimentos.

Aspect-based Sentiment Analysis and Polarity Assignment

ABSTRACT

With the growing expansion of the Web, more and more users share their views on experiences they have had. These views are, in most cases, represented in the form of unstructured text. The Sentiment Analysis (or Opinion Mining) is a research area dedicated to the computational study of the opinions and feelings expressed in texts, typically categorizing them according to their polarity (*i.e.*, as positive or negative). As on-line sales and social networking sites become great sources of opinions, there is a growing need for tools that classify opinions and identify to which aspect of the evaluated entity they refer to. In this work, we propose methods aimed at two key points for the treatment of such opinions: (*i*) aspect-based sentiment analysis and (*ii*) polarity assignment. For aspect-based sentiment analysis, we developed a method that identifies expressions mentioning aspects and entities in text, using natural language processing tools combined with machine learning algorithms. For the identification of polarity, we developed a method that uses 24 attributes extracted from the ranking generated by a search engine to generate machine learning models. Furthermore, the method does not rely on linguistic resources and can be applied to noisy data. Experiments on real datasets show that, in both contributions, our results using a small number of attributes were similar to the baselines. Still, for assigning polarity, the results are comparable to prior art methods that use more complex techniques.

Keywords: opinion mining, sentiment analysis, aspect extraction, text classification.

LISTA DE ABREVIATURAS E SIGLAS

BOW	<i>Bag-of-words</i>
CRF	<i>Conditional Random Fields</i>
EM	<i>Expectation Maximization</i>
HMM	<i>Hidden Markov Models</i>
LDA	<i>Latent Dirichlet Allocation</i>
MNB	<i>Multinomial Naïve Bayes</i>
NB	<i>Naïve Bayes</i>
PLN	Processamento de Linguagem Natural
pLSA	<i>Probabilistic Latent Semantic Analysis</i>
POS	<i>Part-of-Speech</i>
SVM	<i>Support Vector Machine</i>

LISTA DE FIGURAS

Figura 1.1	Transformação do conteúdo gerado por usuário.....	12
Figura 1.2	Visão geral do processo de análise de sentimentos baseada em aspectos	13
Figura 2.1	Exemplo de sumário de opiniões baseado em aspectos.....	17
Figura 2.2	Sumário de opiniões baseado em aspectos	20
Figura 2.3	Visualização de sumários de opiniões baseados em aspectos	21
Figura 2.4	Visão geral de uma Ferramenta de PLN	23
Figura 2.5	Representação gráfica da árvore de relações de dependências	26
Figura 4.1	Visão Geral do Processo	39
Figura 4.2	Fluxo para a categorização de opiniões	40
Figura 4.3	Etapas de transformação de sentenças em instâncias	41
Figura 4.4	Processo de geração de atributos em duas fases	42
Figura 5.1	Visão Geral do Processo	52
Figura 5.2	Sensibilidade de F1 em função de n	62
Figura 6.1	Visualização de sumários de opiniões baseados em aspectos	67
Figura 6.2	Visualização de sumários comparativos de opiniões	68

LISTA DE TABELAS

Tabela 2.1	Lista de POS no modelo PENN TREEBANK PROJECT	24
Tabela 2.2	Lista de relações gramaticais segundo Marnee e Manning (2008)	27
Tabela 3.1	Tipos de atributos utilizados no estado da arte	37
Tabela 4.1	Detalhes dos datasets	45
Tabela 4.2	<i>Baselines</i> para Identificação de Categorias e Alvos	46
Tabela 4.3	Identificação de Categorias.....	47
Tabela 4.4	Identificação do Alvo da Opinião.....	48
Tabela 4.5	Identificação do par (categoria, alvo)	48
Tabela 4.6	Resultados oficiais do SemEval-2015	50
Tabela 5.1	Exemplo de <i>ranking</i> para um <i>post q</i>	55
Tabela 5.2	Atributos derivados do <i>ranking</i>	55
Tabela 5.3	Descrição dos atributos derivados do <i>ranking</i>	56
Tabela 5.4	Detalhes do <i>dataset</i> STD	58
Tabela 5.5	Detalhes do <i>dataset</i> HCR	58
Tabela 5.6	Resultado para atribuição de polaridade em <i>posts</i>	60
Tabela 5.7	Melhores atributos segundo o InfoGain	62

SUMÁRIO

1 INTRODUÇÃO	12
2 REFERENCIAL TEÓRICO	15
2.1 Análise de Sentimentos Baseada em Aspectos	17
2.1.1 Extração da Categoria da Opinião	17
2.1.2 Extração do Alvo da Opinião.....	18
2.1.3 Atribuição da Polaridade do Sentimento	19
2.2 Sumarização de Opiniões	20
2.3 Técnicas e recursos comumente utilizados	22
2.3.1 Ferramentas de PLN	22
2.3.2 Ferramentas de Aprendizagem de Máquina	26
2.4 Métricas de Avaliação	28
2.5 Sumário do Capítulo	29
3 TRABALHOS RELACIONADOS	30
3.1 Análise de Sentimentos Baseada em Aspectos	30
3.2 Atribuição de Polaridade	32
3.3 Sumário dos Trabalhos Relacionados	36
4 EXTRAINDO ENTIDADES E ASPECTOS ATRAVÉS DA CLASSIFICAÇÃO EM DUAS FASES	38
4.1 Visão Geral do Processo	38
4.2 Categorização de opiniões	38
4.2.1 Classificação em uma fase	41
4.2.2 Classificação em duas fases	41
4.3 Identificação do alvo da opinião	42
4.4 Avaliação Experimental	44
4.4.1 Configuração dos Experimentos	44
4.4.2 Metodologia de avaliação e <i>baselines</i>	45
4.4.3 Identificação de Categorias e Alvos da Opinião	47
4.4.4 Análise de Erros	48
4.4.5 Resultados Oficiais SemEval-2015	49
5 ATRIBUIÇÃO DE POLARIDADE BASEADA EM RECUPERAÇÃO DE INFORMAÇÕES	51
5.1 Visão Geral do Processo	51
5.2 Sistemas de Recuperação de Informação	53
5.3 Geração de Atributos	53
5.4 Classificação dos <i>posts</i>	55
5.5 Metodologia de avaliação	56
5.6 Resultados	59
5.6.1 Sensibilidade de Medida-F em função de <i>n</i>	61
5.6.2 Análise dos atributos	61
5.7 Considerações Finais	63
6 SUMARIZAÇÃO DE OPINIÕES	64
6.1 Sumarização de Opiniões sobre Categorias	64
6.2 Sumarização de Opiniões Comparativa	65
7 CONCLUSÃO	69
7.1 Resumo das contribuições	69
7.2 Limitações e trabalhos futuros	70
7.3 Publicações	71
7.4 Outros estudos desenvolvidos	71

REFERÊNCIAS.....	72
-------------------------	-----------

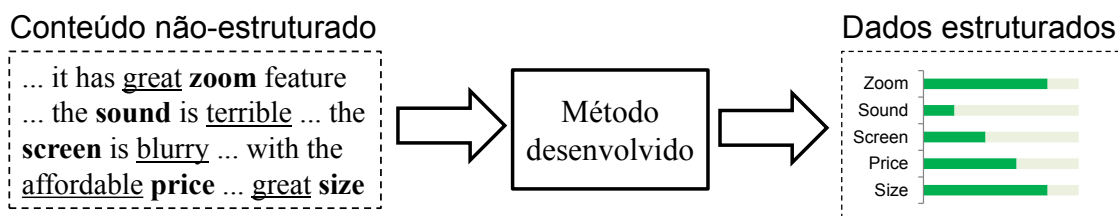
1 INTRODUÇÃO

Atualmente o comportamento dos usuários da Web está mudando, pois além de consumir conteúdos ali disponíveis, eles também estão expondo suas opiniões e experiências, seja sobre um produto que adquiriram, um local que visitaram ou um serviço que utilizaram. Estes relatos podem influenciar na decisão de outros usuários, servindo como informação adicional que frequentemente não está disponível na descrição. Este processo de retroalimentação pode ser bastante útil, tanto para empresas, que podem utilizar destas informações para aprimorar seus produtos, quanto para os consumidores, que também podem tirar proveito de experiência de outros usuários.

A mineração de opinião ou análise de sentimento (*opinion mining/sentiment analysis*) (LIU, 2012) é a área voltada para a análise de opiniões, sentimentos, emoções e atitudes de pessoas e tem se tornado objeto de pesquisa em diversos cenários a partir da mídia social, tais como blogs, e-mails, fóruns de discussão, revisão de produtos, etc. A mineração de opinião faz parte da mineração de textos com foco no processamento de conteúdos gerados por usuários. Esta característica acrescenta diversos desafios de pesquisa tais como identificação de tópicos e de opiniões.

Os conteúdos gerados por usuários são considerados dados não-estruturados, pois podem tratar dos mais diversos assuntos sob a forma de texto livre. A fim de tornar os dados mais facilmente compreensíveis, os métodos desenvolvidos buscam processar as opiniões para que elas possam ser representadas de uma forma estruturada. Dados estruturados permitem uma análise simples e direta sobre os principais tópicos juntamente com a opinião média atribuída. A Figura 1.1 apresenta uma visão geral da transformação de um conteúdo não-estruturado (textual) para uma representação estruturada (sumário).

Figura 1.1: Transformação do conteúdo gerado por usuário

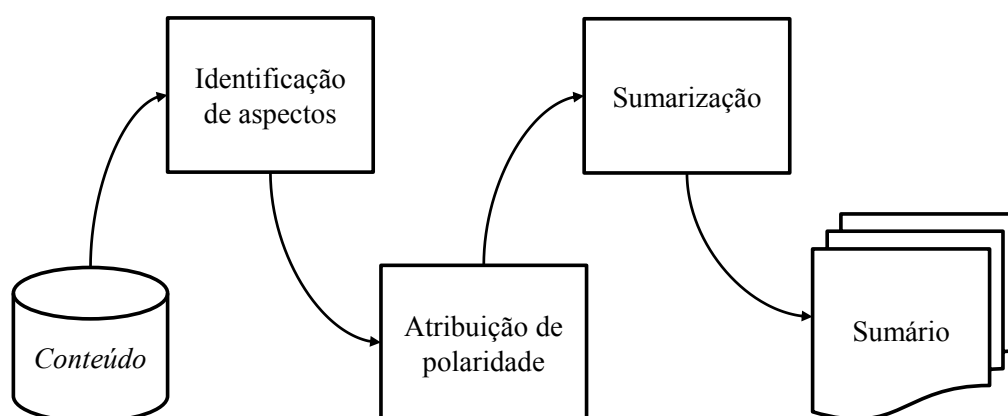


Fonte: O autor

Com a crescente popularização das mídias sociais, ferramentas capazes processar opiniões têm que lidar com grandes quantidades de dados. Assim, torna-se necessária a

representação dos dados de alguma forma sumarizada. A análise de sentimentos baseada em aspectos é o segmento da mineração de opinião que busca identificar os aspectos (tópicos, assuntos, etc.) mencionados em um texto e determinar a polaridade do sentimento (positivo e negativo). O processo consiste de três etapas fundamentais: (i) identificação dos aspectos mencionados; (ii) atribuição de polaridade; e (iii) sumarização. A Figura 1.2 apresenta a visão geral deste processo.

Figura 1.2: Visão geral do processo de análise de sentimentos baseada em aspectos



Fonte: O autor

A motivação deste trabalho dá-se pela crescente demanda por ferramentas que processem opiniões, no meio acadêmico quanto no meio corporativo no qual empresas precisam avaliar as opiniões sobre seus produtos. Ao mesmo tempo, ferramentas recentes tem se tornando cada vez mais complexas devido ao uso extensivo de recursos linguísticos que agregam alto custo de processamento. Outros dois fatores motivam este trabalho: a dependência de dados anotados manualmente e o número praticamente ilimitado de maneiras que as pessoas podem expressar suas opiniões (LIU, 2012).

O objetivo deste trabalho é contribuir para a análise de sentimentos, desenvolvendo métodos que realizem (i) a identificação de aspectos e (ii) a atribuição de polaridade em conteúdos gerados por usuários. Os métodos desenvolvidos em ambas as contribuições visam reduzir o número de atributos utilizados para a classificação de textos, obtendo resultados similares aos dos *baselines*.

Especificamente, para identificação de aspectos, desenvolvemos um método que identifica expressões que mencionem aspectos e entidades em um texto, utilizando ferramentas de processamento de linguagem natural combinadas a algoritmos de aprendizagem de máquina. O método desenvolvido foi avaliado no âmbito da campanha SemE-

val 2015¹, tendo sido classificado entre o 7º e o 13º lugar, utilizando somente 20 atributos.

O método para a atribuição de polaridade baseou-se na hipótese de que conteúdos similares tendem a compartilhar a mesma polaridade de sentimento. Desta forma, o texto a ser classificado é tratado como uma consulta e submetido a um motor de busca que indexa textos cujas classes de sentimento são conhecidas. O motor de busca retorna o ranking dos textos mais similares à consulta. Com base nesse ranking são calculados 24 atributos que representam informações sobre a polaridade do sentimento. O método não depende de recursos linguísticos e pode ser aplicado sobre dados com ruídos. Os experimentos foram realizados sobre *datasets* reais demonstram os resultados são comparáveis aos de métodos do estado da arte que utilizam técnicas mais complexas.

O restante deste trabalho está dividido na seguinte forma: no Capítulo 2 são introduzidos os fundamentos básicos necessários para o entendimento deste trabalho, além de uma revisão de conceitos utilizados e detalhes de ferramentas e métodos amplamente utilizados na literatura e que fazem parte dos métodos desenvolvidos. No Capítulo 3 são discutidos os trabalhos relacionados. No Capítulo 4 é descrito o método desenvolvido para análise de sentimentos baseada em aspectos. No Capítulo 5 é descrito o método desenvolvido para atribuição de polaridade. No Capítulo 6 são apresentados os sumários de aspectos para os métodos desenvolvidos. Finalmente, o Capítulo 7 conclui este trabalho e apresenta algumas direções para trabalhos futuros.

¹<<http://alt.qcri.org/semEval2015/task12/>>

2 REFERENCIAL TEÓRICO

Neste capítulo serão abordadas algumas definições usadas neste trabalho que são necessárias para o entendimento das atividades realizadas.

A análise de sentimentos, também chamada de mineração de opinião, é a área de estudo que analisa as opiniões que as pessoas expressam sobre um determinado assunto (LIU, 2012). Uma opinião pode representar um sentimento, avaliação, atitude, e emoção. Com a popularização de mídias sociais (por exemplo redes sociais, fóruns de discussão, blogs, etc.), ferramentas que identificam e processam opiniões são de grande importância no meio corporativo (por exemplo, opiniões sobre produtos ou serviços). O objetivo é fornecer aos usuários informações relevantes sobre experiências de outros usuários de maneira objetiva, sem a necessidade de visualizar uma grande quantidade dados.

Sistemas que processam opiniões automaticamente são uma alternativa viável quando comparados a métodos de pesquisas diretamente com os consumidores ou processamento manual destas opiniões. Além disso, fornecem informações importantes para a análise de aceitação (de um produto, serviço, candidato político, etc.) sem a necessidade do preenchimento de formulários por parte do usuário. A mineração de opinião concentra-se em analisar opiniões que expressam ou implicam sentimentos positivos e negativos, embora em alguns casos seja interessante também processar as informações neutras (LIU, 2012).

A opinião também pode ser processada em diferentes níveis de granularidade, como:

- *Nível de documento* entende que cada documento expressa uma opinião geral (positiva ou negativa) sobre uma entidade (e.g. produto, serviço, evento, etc.);
- *Nível de sentença* entende que cada sentença (e.g. frase ou cláusula) expressa uma opinião a respeito de um ou vários aspectos (e.g. preço, qualidade, etc.) de uma entidade;
- *Níveis de entidades e aspectos* considera que cada opinião consiste de um sentimento (positivo ou negativo) e um alvo (da opinião). Um documento pode conter opiniões sobre diversos aspectos de diversas entidades. Nesta granularidade é obtido o máximo de detalhes sobre o que o usuário gosta e o que não gosta.

Na mineração de opinião, os principais indicadores de sentimentos são as palavras chamadas *opinion words* ou palavras de opinião. Estas palavras estão associadas a sentimentos, por exemplo *good*, *wonderful*, e *amazing* são palavras positivas, enquanto

que *bad*, *poor*, e *terrible* são palavras negativas. Uma lista de palavras de opinião é chamada de *opinion lexicon* ou dicionário de opiniões. Entretanto, um dicionário pode ser insuficiente para a análise de opinião, principalmente porque algumas palavras podem ter sentimentos opostos em domínios diferentes; algumas sentenças não expressam opiniões (e.g. sentenças interrogativas, condicionais, comparativas, etc.); e sentenças irônicas e/ou sarcásticas representam ruído nos dados.

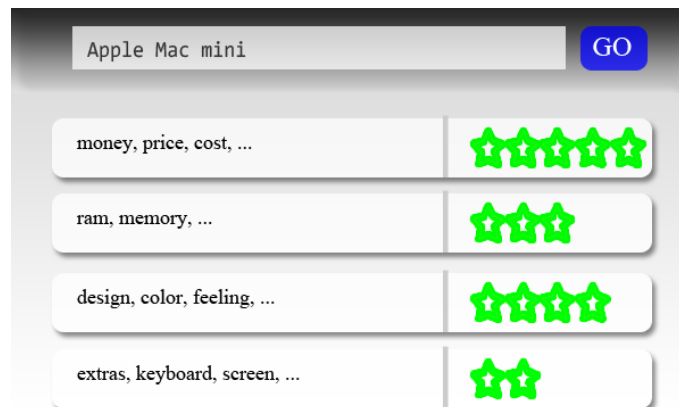
A mineração de opinião pode analisar uma variada gama de documentos: notícias, *tweets* (postagem de Twitter), fóruns de discussão, blogs, revisões, etc. Neste trabalho, utilizaremos o termo *post* para representar um texto (formal ou informal) que expressa opiniões. Uma opinião contém dois componentes principais: um alvo g e um sentimento s sobre o alvo. Formalmente, uma opinião é um par (g, s) onde g é a expressão que se refere a uma entidade ou aspecto de uma entidade e s representa o sentimento atribuído (positivo, negativo ou neutro) ou a pontuação numérica que representa a intensidade do sentimento.

Por exemplo, considerando a sentença "*The phone has a good screen*", como saída é obtida a opinião $(screen, good)$, que pode ser generalizada para $(screen, positivo)$. Neste trabalho, um *dataset* é um conjunto de *posts* que pertencem ao mesmo domínio ou origem e que contém rótulos dos dados (e.g. opiniões, entidades, aspectos, alvos ou sentimentos). Esses rótulos representam as classes que um sistema de análise de sentimentos busca identificar.

A classificação de opinião em nível de documento geralmente é insuficiente, pois avalia a opinião em alto nível para uma entidade desconsiderando seus aspectos individualmente. Além disso, uma opinião pode ao mesmo tempo avaliar diversos aspectos de uma entidade. Considerar uma opinião em nível de documento como positiva, implica que todos os aspectos relacionados a uma entidade são positivos, o que pode não ser verdade na prática (LIU, 2012). Assim, para uma avaliação mais precisa, é necessário identificar todos os aspectos e respectivas opiniões sobre uma entidade.

A análise de sentimentos em nível de entidade e aspectos permite estruturar sentimentos produzindo sumários que representam a intensidade do sentimento sobre cada aspecto, transformando textos não estruturados em dados estruturados que facilitam o entendimento. A Figura 2.1 apresenta um exemplo de sumário de opiniões baseado em aspectos para a entidade *Apple Mac mini*. O exemplo apresenta cada aspecto (e.g. preço, *hardware*, *design*, etc.) e a respectiva intensidade do sentimento atribuído.

Figura 2.1: Exemplo de sumário de opiniões baseado em aspectos



Fonte: Pontiki et al. (2015)

2.1 Análise de Sentimentos Baseada em Aspectos

A Análise de Sentimentos Baseada em Aspectos busca obter o máximo de detalhes sobre todas as entidades e aspectos mencionados em um *post*. Neste trabalho, a hierarquia das informações em um *dataset* obedece à seguinte ordem: um domínio é a fonte (ou origem) limitada de conhecimento; um *dataset* contém um conjunto de *posts* de um mesmo domínio; um *post* contém uma ou mais sentenças; uma sentença contém zero ou mais opiniões.

Na análise de sentimentos baseada em aspectos, uma opinião O é composta por uma tripla $O = (C, T, S)$, onde C é uma categoria que representa o par $C = (E, A)$, sendo E uma entidade e A é um aspecto da entidade E ; T é o alvo da opinião mencionado explicitamente no *post* ou "NULL" quando mencionado implicitamente; S é o sentimento atribuído à opinião, pertencente ao conjunto $\{positivo, negativo, neutro\}$. Cada domínio contém uma lista própria de entidades possíveis, outrossim, cada entidade contém uma lista própria de aspectos. Para a identificação de cada um dos três componentes uma tarefa separada é necessária. Estas serão descritas a seguir.

2.1.1 Extração da Categoria da Opinião

Cada categoria C representa um par único em um determinado domínio. Isso faz com que o uso de dicionários seja insuficiente para identificar sinônimos em diferentes

domínios, principalmente pelo fato das pessoas usarem diferentes palavras para descrever uma mesma entidade (LIU, 2012).

Por exemplo, considere a seguinte sentença: "*The voice quality of this phone is amazing.*" a categoria é {PHONE, VOICE_QUALITY}, sendo VOICE_QUALITY o aspecto da entidade PHONE. Neste exemplo, as informações quanto a categoria estão explícitas no texto, tanto a entidade quanto ao aspecto.

Como um segundo exemplo, considere a seguinte sentença: "*The applications are also very easy to find and maneuver.*" a categoria é {SOFTWARE, USABILITY}, sendo USABILITY o aspecto da entidade SOFTWARE. Neste exemplo, nota-se que tanto a entidade SOFTWARE quanto o aspecto USABILITY não estão explícitos no texto, tratando-se de entidades e aspectos implícitos. Assim, uma categoria pode ser considerada uma informação implícita da opinião.

A extração de categorias (também chamada de categorização de opiniões) é difícil devido à necessidade de grande quantidade de conhecimento *a priori*. As principais estratégias utilizam combinações de métodos supervisionados, não-supervisionados e regras de associação (LIU, 2012). Abordagens para a extração de categorias são detalhadas na Seção 3.1.

2.1.2 Extração do Alvo da Opinião

O alvo da opinião T é uma expressão usada em um *post* para se referir a uma entidade E do par (E, A) . O alvo é definido por sua posição única de início e final em um texto. O alvo pode não estar explícito no texto (assumindo o valor "NULL"), por exemplo, usando pronomes ou correferências textuais. Para a identificação de alvos explícitos existem quatro formas principais:

1. **Extração baseada em substantivos frequentes:** Utiliza analisador gramatical para identificar a substantivos mais frequentes. Esta abordagem foi desenvolvida inicialmente por Hu e Liu (2004);
2. **Extração através da relação entre o alvo e o sentimento:** Utiliza analisador gramatical e relações de dependência para obter evidências de relação entre o alvo e as palavras de sentimento. Um trabalho bastante detalhado foi proposto por Qiu et al. (2011);
3. **Extração usando aprendizado supervisionado:** Utiliza modelos de aprendizado

supervisionado para determinar se uma opinião refere-se a uma entidade ou aspecto da entidade. Diversos trabalhos utilizam esta abordagem sendo mais comum para identificação de entidades implícitas (PONTIKI et al., 2015);

4. **Extração usando modelos de tópicos:** Utiliza métodos baseados em agrupamentos (ou *cluster*) de tópicos buscando obter distribuições que representem aspectos. Diversas abordagens tem sido apresentadas na literatura tais como utilização de pLSA (Análise Semântica Latente Probabilística) (HOFMANN, 1999), LDA (Alocação Latente de Dirichlet) (BLEI; NG; JORDAN, 2003), EM (Máxima Expectativa) (DEMPSTER; LAIRD; RUBIN, 1977), entre outros.

Considerando a sentença do exemplo anterior: "*The applications are also very easy to find and maneuver.*" o alvo da opinião é "*applications*", que em um domínio de laptops refere-se à entidade SOFTWARE. Além disso, a expressão "*very easy to find and maneuver*" refere-se ao aspecto USABILITY. As principais abordagens utilizadas para esta tarefa consistem de *n*-gramas, POS tags, prefixos, sufixos, informações externas provenientes modelos de tópicos, etc. Outras abordagens para a extração do alvo da opinião são discutidas na Seção 3.1.

2.1.3 Atribuição da Polaridade do Sentimento

Cada opinião identificada em um *post* tem uma polaridade atribuída a partir do conjunto $P = \{\textit{positivo}, \textit{negativo}, \textit{neutro}\}$. A classificação de um sentimento neutro ocorre quando não há uma definição clara sobre sua polaridade (positiva ou negativa). Para a determinar a polaridade de cada opinião, existem duas formas principais:

- **Atribuição baseada em aprendizado supervisionado:** utiliza aprendizado supervisionado baseado em nível de sentença para determinar a polaridade de uma opinião. Neste caso, a sentença pode representar o escopo (i.e. palavras que se relacionam com o aspecto de interesse) da expressão de sentimento referentes à opinião. Esta abordagem torna o método dependente dos dados de treinamento, fazendo com que o modelo treinado obtenha piores resultados quando aplicado a domínios diferentes (LIU, 2012);
- **Atribuição baseada em informações léxicas:** são métodos geralmente supervisionados que utilizam dicionários de opiniões e recursos de processamento como analisadores gramaticais, árvores de dependências, etc. para determinar a polari-

dade de uma opinião. Estes recursos serão detalhados na seção 2.3

Considerando a mesma sentença: "*The applications are also very easy to find and maneuver.*" a polaridade da opinião é *positivo*. Esta evidência está diretamente associada à palavra de opinião "*easy*" que ainda recebe o um modificador de intensidade "*very*". Finalmente, após as três tarefas, é extraída a opinião O da sentença como sendo: $O = (\{\text{SOFTWARE, USABILITY}\}, \text{"applications"}, \text{positivo})$. Abordagens para a atribuição de polaridade são reportadas na Seção 3.2.

2.2 Sumarização de Opiniões

A sumarização de opiniões permite uma representação simples e direta a respeito de algum aspecto da entidade, juntamente com a avaliação quantitativa representada pelo número de pessoas que a avaliaram de forma positiva e negativa aquele aspecto. Esta representação, é chamada de sumário de opiniões baseado em aspectos. A Figura 2.2 mostra um exemplo de sumário de opiniões sobre uma câmera digital (HU; LIU, 2004). Considerando a câmera digital como entidade, o aspecto GENERAL representa opiniões que se referem diretamente a ela. Para cada aspecto (ou componente), por exemplo, *picture quality* é mostrado o número de pessoas que o consideraram como positivo e negativo.

Figura 2.2: Sumário de opiniões baseado em aspectos

Digital Camera 1:

Aspect: GENERAL	
Positive:	105
Negative:	12
Aspect: Picture quality	
Positive:	95
Negative:	10
Aspect: Battery life	
Positive:	50
Negative:	9
...	

Fonte: Liu (2012)

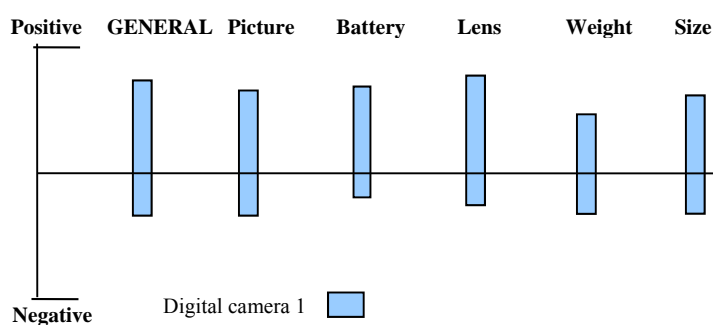
A Figura 2.3a utiliza um gráfico de colunas para mostrar o sumário da Figura 2.2. Na figura, para cada aspecto identificado no topo, a parte da coluna que está acima do eixo x mostra o número de opiniões positivas. As opiniões negativas estão abaixo do eixo x . Esta visualização também pode ser convertida para produzir o gráfico apresentado na

Figura 2.1. Para tal, basta dividirmos o número de opiniões positivas pelo número total de opiniões sobre o aspecto.

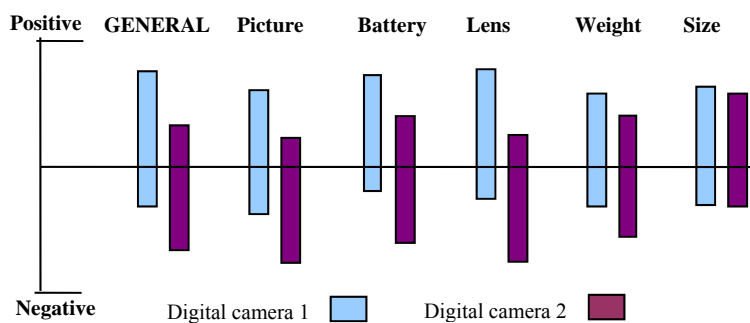
A representação de aspectos através de sumários estruturados também permite a comparação entre produtos de uma mesma categoria, e.g. câmeras, um exemplo desta visualização é apresentado na Figura 2.3b. A visualização comparativa de sumários foi objeto de estudo no trabalho de Liu, Hu e Cheng (2005).

Figura 2.3: Visualização de sumários de opiniões baseados em aspectos

(a) Visualização de um sumário de opiniões sobre uma câmera digital



(b) Comparação visual de duas câmeras digitais



Fonte: Liu (2012)

A sumarização de opiniões (LIU, 2012; HU; LIU, 2004; LIU; HU; CHENG, 2005) é uma técnica bastante útil para análise de opiniões de um grande número de usuários. Assim, um sumário poderia conter as principais características sobre determinados aspectos bem como uma escala de aceitação de um determinado produto (LIU, 2012). A sumarização é a etapa final da análise de sentimentos baseada em aspectos. Como o foco deste trabalho não é sumarização, neste trabalho apenas aplicamos métodos existentes para podermos mostrar o resultado do processo de análise de sentimentos.

2.3 Técnicas e recursos comumente utilizados

Ferramentas utilizadas para o processamento de *posts* têm origem na área de Processamento de Linguagem Natural (PLN) Estas ferramentas tratam a forma de representação do conteúdo de um *post*, e geralmente adicionam alto custo de processamento aos métodos de análise de sentimentos. Um *post* é um texto que foi escrito por um usuário da Web, tratando-se muitas vezes de textos informais. O tratamento de um *post* pode necessitar de diversos tipos de processamento para que se possa obter a informação desejada. Nesta seção serão apresentadas as principais técnicas de processamento e recursos que são comumente utilizados na literatura.

2.3.1 Ferramentas de PLN

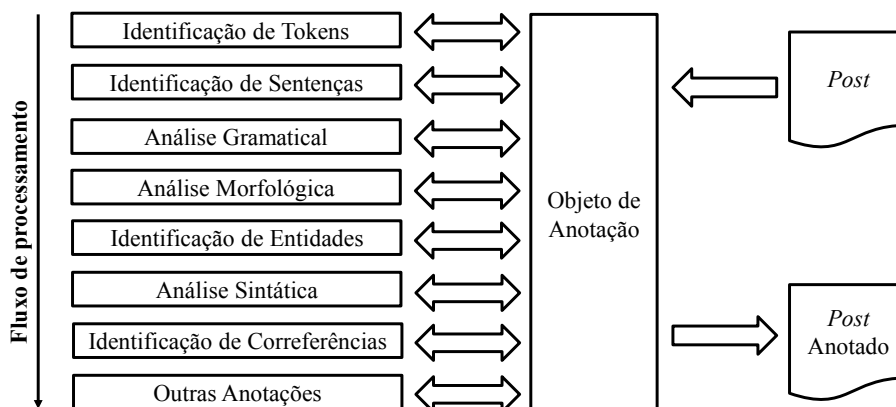
Ferramentas de PLN, são utilizadas para adicionar informações não explícitas a um *post*. Estas informações são chamadas de **anotações**¹ e requerem algum esforço computacional, dependendo da sua complexidade. Atualmente existem diversas ferramentas que oferecem um *framework* completo para o processamento de textos, tais como GATE (CUNNINGHAM et al., 2002), NLTK (LOPER; BIRD, 2002), UIMA (FERRUCCI; LALLY, 2004), e Stanford Core NLP Toolkit (MANNING et al., 2014). Em comum entre elas está o objetivo de oferecer os principais métodos para o processamento de textos.

A Figura 2.4 apresenta a visão geral de uma ferramenta de PLN. Como entrada, o sistema recebe um texto, que no contexto deste trabalho trabalho é representado por um *post*. Internamente, um *post* é representado por um **Objeto de Anotação** que é submetido a diferentes recursos de processamento. Estes recursos, são chamados de *recursos externos* pois são informações que não estão explícitas no texto e agregam elevado custo computacional para sua execução. Como saída a ferramenta produz um conjunto de anotações que combinadas ao *post* original geram um ***post anotado***.

Análise Léxica. Realiza o particionamento do *post* em *tokens* (símbolos) que são geralmente palavras. Esta etapa pode ser considerada a mais simples, sendo que alguns modelos consideram o caractere de espaço como separador, gerando uma sequencia de

¹Neste trabalho, o termo anotação é usado para referir-se às etiquetas agregadas ao texto pelas ferramentas de PLN. Utilizaremos o termo rótulo para referir-se às classes que o sistema de análise de sentimento deve identificar.

Figura 2.4: Visão geral de uma Ferramenta de PLN



Fonte: Adaptado de Manning et al. (2014)

tokens. No entanto, para alguns idiomas onde o espaço não significa um delimitador de palavras (por exemplo, chinês) devem ser utilizadas outras abordagens tais como expressões regulares, dicionários, etc. Além disso, *tokens* devem ser únicos em um *post*, portanto, é necessário identificar-los através de sua posição única no texto.

Identificação de sentenças. Identifica as sentenças de um *post*, separando conjuntos de *tokens* que pertençam a mesma sentença. Sentenças geralmente são simples de ser identificadas, por exemplo, considerando caracteres de pontuação (.!?) como delimitadores. No entanto, existem algumas dificuldades quanto a essa abordagem, tais como abreviaturas, números, etc. Nestes casos, são analisados os caracteres próximos aos caracteres de pontuação. Por exemplo, se a próxima letra é maiúscula é provável que represente o final de uma sentença, se o caractere anterior for um número e o próximo também, provavelmente trata-se de uma representação numérica.

Análise Gramatical. Realizada um processo chamado de *part-of-speech tagger* (POS *tagger*), (etiquetador morfológico) que determina a classe gramatical de cada *token* (substantivos, verbos, adjetivos, etc). Geralmente este recurso utiliza a combinação de dicionários e modelos de aprendizagem (supervisionados e não supervisionados), que são recursos com elevado custo computacional. Uma abordagem bastante difundida na literatura foi desenvolvida por Santorini (1990) chamada *Penn Treebank Project* que utiliza um método baseado em regras. Mais recentemente, Toutanova et al. (2003) desenvolveu um método que atribui a classe gramatical baseada em um modelo de máxima entropia. A Tabela 2.1 apresenta um conjunto padrão de POS *tags*.

Análise Morfológica e Linguística. Executa tarefas de transformação do texto, tais como: correção ortográfica, *truecasing*, lematização, etc. A correção ortográfica pode

Tabela 2.1: Lista de POS no modelo PENN TREEBANK PROJECT

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

Fonte: Adaptado de Santorini (1990)

ser aplicada de diferentes formas, no entanto, os métodos mais comuns são baseados em distância de edição, também conhecida como *distância de Levenshtein*; e baseados em interseção de segmentos do *token*, por exemplo, o *coeficiente de Jaccard* (MANNING; RAGHAVAN; SCHÜTZE, 2008). *Truecasing* (LITA et al., 2003) é o processo que visa restaurar a forma original e correta dos caracteres de *tokens*, por exemplo, erros de escrita de palavras em maiúsculo e minúsculo. As principais abordagens utilizam métodos estatísticos e modelos de linguagem (LITA et al., 2003) e modelos baseados em *Conditional Random Fields* (CRF) (FINKEL; GRENAGER; MANNING, 2005). A lematização visa identificar a forma canônica de cada *token*, chamada de lema. A representação de *tokens* através de lemas reduz a quantidade de derivações de uma mesma palavra e necessitam de amplo conhecimento do vocabulário e análise morfológica completa do idioma (MANNING; RAGHAVAN; SCHÜTZE, 2008).

Identificação de Entidades Nomeadas. Atribui o tipo de entidade representado por um *token* pertencente ao conjunto {PESSOA, LOCAL, ORGANIZAÇÃO, OUTROS} e para números pertencentes ao conjunto {DINHEIRO, NÚMERO, DATA, HORA, DURAÇÃO} (MANNING et al., 2014). Entidades são reconhecidas combinando modelos baseados em CRF (FINKEL; GRENAGER; MANNING, 2005) para entidades, baseados

em regras (para números) e baseados em expressões temporais (CHANG; MANNING, 2012).

Análise Sintática. Também chamada de *parsing*, analisa a sequência de *tokens* para determinar a estrutura gramatical. A estrutura gramatical consiste na representação das relações de dependência entre *tokens*, geralmente através de uma árvore. Na da análise léxica são obtidos o *tokens*, para que o analisador sintático utilize modelos probabilísticos para construir uma árvore sintática da estrutura (KLEIN; MANNING, 2003; CHEN; MANNING, 2014). A Figura 2.5 mostra a representação gráfica da árvore de relações de dependências para a sentença "*Bills on ports and immigration were submitted by Senator Brownback, Republican of Kansas*" (MARNEE; MANNING, 2008). A Figura 2.5a mostra as relações de dependências básicas enquanto que a Figura 2.5b mostra as relações de dependências simplificadas. A Tabela 2.2 apresenta o conjunto de relações segundo Marnee e Manning (2008). O conjunto de relações é composto por aproximadamente 50 relações gramaticais. Todas as dependências são relações binárias, tendo um **governante** e um **dependente**. O exemplo da Figura 2.5 são identificadas as seguintes relações:

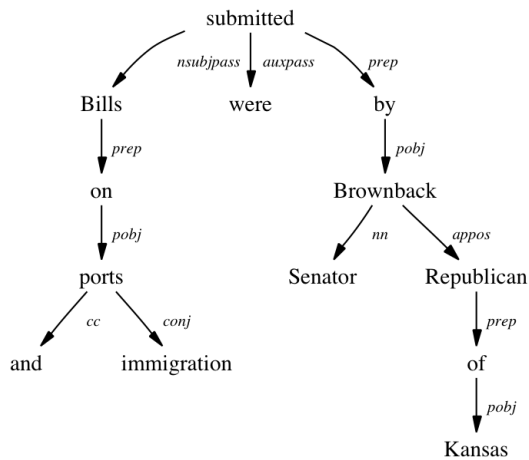
nsubj (makes-8, Bell-1)	amod (products-16, electronic-11)
nsubj (distributes-10, Bell-1)	conj_and (electronic-11, computer-13)
vmod (Bell-1, based-3)	amod (products-16, computer-13)
nn (Angeles-6, Los-5)	conj_and (electronic-11, building-15)
prep_in (based-3, Angeles-6)	amod (products-16, building-15)
root (ROOT-0, makes-8)	dobj (makes-8, products-16)
conj_and (makes-8, distributes-10)	dobj (distributes-10, products-16)

Identificação de Correferências. Também chamada de resolução de correferências, busca encontrar todas as expressões que se referem a mesma entidade em um *post*. Lee et al. (2013) propõem uma abordagem que combina o uso de modelos de aprendizagem supervisionados e não-supervisionados, utilizando um sistema baseado em regras chamado *entity-centric coreference*. Este método foi utilizado como parte do trabalho desenvolvido para identificação de correferências nas etapas de identificação de categorias e extração do alvo da opinião.

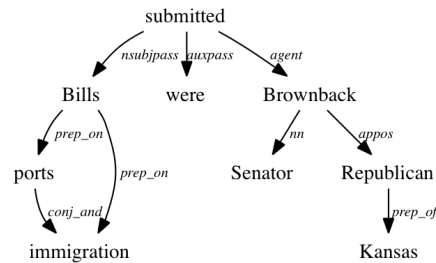
Outras Anotações. Adicionam informações subjetivas ao *post*, tais como, gênero, emoção, sentimento, etc.

Figura 2.5: Representação gráfica da árvore de relações de dependências

(a) Relações de dependências básicas



(b) Relações de dependências simplificadas



Fonte: Manning et al. (2014)

2.3.2 Ferramentas de Aprendizagem de Máquina

A aprendizagem de máquina explora o estudo e construção de algoritmos que podem aprender e fazer previsões sobre os dados. Algoritmos de aprendizagem produzem modelos a partir de exemplos de entrada (também chamados de instâncias), usados como treinamento. Os modelos são conjuntos de hipóteses formuladas sobre os dados de entrada para realizar a previsão de dados não vistos no treinamento. Cada instância é representada por um vetor de atributos. Os atributos são informações que caracterizam uma instância. A classe (ou rótulo) é o atributo no qual o modelo de aprendizagem faz a previsão baseado nos outros atributos. Ao atribuir a classe a uma instância não-rotulada realiza-se a classificação.

Modelos de aprendizagem são de grande importância para a análise de sentimentos, visto que são responsáveis por agregar inteligência aos métodos (LIU, 2012). Atualmente existem diversas ferramentas que oferecem um conjunto métodos para o processamento de dados. Weka (HALL et al., 2009), por exemplo, apresenta um conjunto de algoritmos de aprendizagem de máquinas voltados para a mineração de dados. Além disso, também oferece os principais métodos utilizados no decorrer deste trabalho, tais como pré-processamento, classificação, seleção de atributos, etc.

Pré-processamento. Realiza a representação dos atributos das instâncias. Na análise de sentimentos, os atributos geralmente consistem de *tokens* e podem ser obtidos da mesma

Tabela 2.2: Lista de relações gramaticais segundo Marnee e Manning (2008)

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	punct	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
mwe	Expressão multi-palavra	xcomp	Complemento clausal aberto
neg	Modificador de negação	xsubj	Sujeito controlador
nn	Substantivo composto		

Fonte: O autor

forma como em ferramentas de PLN. Os atributos são específicos de um *dataset*, assim, cada instância contém todos os *tokens* possíveis como atributos. O valor de um atributo pode indicar a presença do *token* no *post*, bem como a frequência no *post* ou no *dataset*. A utilização de *tokens* isoladamente é conhecida como *bag-of-words* (BOW), e não leva em conta a ordem dos *tokens* em um *post*. Para preservar a ordem dos *tokens* utiliza-se uma técnica chamada de *n-gramas*. O valor de *n* indica a quantidade de *tokens* considerada na sequência. Por exemplo, *unigramas* considera que cada atributo é representado por um *token*, *bigramas* considera a sequência de dois *tokens*, geralmente utilizada para identificar expressões.

Modelos de aprendizagem supervisionada. São modelos que identificam a relação dos atributos com o rótulo da instância. Na análise de sentimentos, a atribuição de polaridade é tipicamente tratada como uma tarefa de classificação (LIU, 2012). Assim, dado um conjunto de atributos que constituem um *post* é possível atribuir um rótulo (polari-

dade) através da classificação supervisionada. A classificação supervisionada também é importante para outras tarefas, como a identificação de categorias da opinião e extração do alvo. Os algoritmos de classificação mais comuns para a análise de sentimentos são Naïve Bayes (NB), Máquina de Vetor de Suporte (SVM), Máxima Entropia (MaxEnt), etc.

Seleção de atributos. A seleção de atributos consiste em analisar os atributos mais importantes para a determinação do rótulo. Esta análise é importante porque serve como um filtro que remove *tokens* menos discriminantes, reduzindo o número de atributos no modelo. Além disso, serve para ordenar os atributos conforme sua importância. Neste trabalho, foi utilizada a seleção de atributos como parte do método desenvolvido. A técnica de seleção utilizada é `InfoGainAttributeEval` (MITCHELL, 1997), esta técnica também exige um método de busca que atribui um escore ao atributo, o método utilizado foi o `Ranker`, disponível no próprio Weka (HALL et al., 2009). Com a combinação destes métodos são obtidos os atributos utilizados para a identificação de categorias e as palavras candidatas para identificação do alvo da opinião (mais detalhes no Capítulo 4).

Modelos de aprendizagem não-supervisionada. Modelos não-supervisionados referem-se a classe de problemas onde o objetivo é determinar como os dados estão organizados. Frequentemente considera como as informações estão agrupadas, criando agrupamentos (*clusters*) de instâncias próximas. Estes agrupamentos não consideram o rótulo, mas as informações latentes, e geralmente seguem distribuições probabilísticas. Métodos não supervisionados são frequentemente utilizados no estado da arte (QIU et al., 2011; LIU, 2012; LIU et al., 2013). Métodos baseados em regras também são considerados como não-supervisionados. No método desenvolvido são utilizadas algumas regras para a identificação do alvo da opinião (mais detalhes no Capítulo 4).

2.4 Métricas de Avaliação

Para a avaliação dos resultados de algoritmos de atribuição de polaridade e de identificação de aspectos, há um conjunto de métricas padrão: acurácia (Acc), precisão (Pr), revocação (Rec) e Medida-F (F1). Essas métricas comparam os resultados obtidos contra os resultados esperados (*i.e.*, determinados pelos rótulos das classes) e são definidas como:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2.1)$$

$$Pr = \frac{TP}{TP + FP}, \quad (2.2)$$

$$Rec = \frac{TP}{TP + FN}, \quad (2.3)$$

$$F1 = 2 \times \frac{Pr \times Rec}{Pr + Rec} \quad (2.4)$$

onde TP, TN, FP, e FN referem-se a verdadeiro positivo, verdadeiro negativo, falso positivo, e falso negativo, respectivamente. Pr, Rec e F1 são micro-médias, nas quais as instâncias têm pesos iguais.

2.5 Sumário do Capítulo

Neste capítulo foram definidos os conceitos relacionados à análise de sentimentos bem como as principais tarefas e ferramentas utilizadas neste trabalho. A análise de sentimentos baseada em aspectos oferece a forma mais completa de análise opiniões expressas por usuários em *posts*. No entanto, trata-se de uma tarefa bastante complexa que envolve técnicas de diversas áreas de conhecimento, tais como mineração de dados, processamento de linguagem natural e aprendizagem de máquina.

Os fundamentos descritos neste Capítulo ajudam a entender o funcionamento de métodos do estado da arte e do trabalho desenvolvido. No próximo capítulo, serão apresentados os trabalhos do estado da arte e suas relações com o trabalho desenvolvido.

3 TRABALHOS RELACIONADOS

Neste capítulo serão apresentados os principais trabalhos relacionados ao tema dessa dissertação. Serão discutidos os trabalhos da literatura que tratam dos dois pontos de contribuição deste trabalho: a análise de sentimentos baseada em aspectos e a atribuição de polaridade. Assim, as próximas Seções deste Capítulo estão organizadas da seguinte forma: na Seção 3.1 serão abordados os trabalhos voltados à análise de sentimentos que tratam as tarefas de extração e categorização de aspectos; na Seção 3.2 serão abordados os trabalhos que tratam da atribuição de polaridade em *posts*.

3.1 Análise de Sentimentos Baseada em Aspectos

A Análise de Sentimentos Baseada em Aspectos busca identificar as opiniões expressas sobre todos os aspectos e entidades mencionados em um *post* (LIU, 2012). Para isso, é necessário identificar e extrair as palavras que representam estes aspectos e entidades. Para a extração de aspectos existem quatro abordagens principais:

- extração baseada em substantivos frequentes;
- extração através da relação entre o alvo e o sentimento;
- extração usando aprendizagem supervisionada; e
- extração usando modelos de tópicos.

A *extração baseada em substantivos frequentes* tem como objetivo encontrar expressões com aspectos explícitos, representados por substantivos a partir de grandes *datasets*. Hu e Liu (2004) utilizaram um algoritmo de mineração de dados explorando *POSTags*. Esta abordagem é bastante interessante porque supõe que usuários tendem a utilizar o mesmo vocabulário para definir um mesmo aspecto, enquanto que conteúdo irrelevante é geralmente diferente nos diversos comentários. A utilização de coocorrências de palavras foi estudada por Moghaddam e Ester (2010) que buscaram identificar os aspectos mais frequentes e remover os substantivos irrelevantes. A remoção baseou-se no número de ocorrências no *dataset*, utilizando remoção de afixos e remoção de palavras comuns. A extração baseada em substantivos frequentes é a forma mais simples de extração de aspectos seus resultados tem servido como *baseline* para métodos mais complexos. A desvantagem dessa abordagem é que muitos erros de extração ocorrem devido a substantivos frequentes que não representam um aspecto real (QIU et al., 2011).

A *extração através da relação entre o alvo e o sentimento* utiliza palavras de sentimentos frequentes para identificar aspectos pouco frequentes. Considerando que uma mesma palavra de opinião pode ser utilizada para caracterizar diferentes aspectos, as relações sintáticas entre palavras de opinião e alvos ajudam a extrair aspectos. Da mesma forma, substantivos frequentes ajudam a identificar palavras de opinião pouco frequentes. A extração de aspectos a partir de relações sintáticas foi inicialmente utilizada por Qiu et al. (2011). Neste trabalho, os autores utilizam uma gramática de dependência para descrever as relações sintáticas, considerando que as relações entre aspectos e opiniões ocorrem somente de forma direta, que aspectos podem ser somente substantivos e que opiniões são somente representadas por adjetivos. De forma semelhante, Liu et al. (2013) utilizaram-se destes conceitos para aperfeiçoar o método ao remover palavras que não podem ser aspectos, desta forma, melhorando a precisão do método e sem perda significativa da revocação. A principal desvantagem desta abordagem deve-se ao fato de que alguns aspectos são expressões multi-palavras e muitas vezes estes aspectos estão implícitos. Entretanto, o método é bastante eficaz principalmente por identificar palavras de opinião independentemente do domínio.

A *extração usando aprendizagem supervisionada* trata a extração de aspectos como uma tarefa classificação. Assim, estas técnicas exigem grandes *datasets* rotulados manualmente para treinamento (LIU, 2012). As principais técnicas de extração de aspectos supervisionadas utilizam *Hidden Markov Models*(HMM) (RABINER, 1989) e *Conditional Random Fields* (CRF) (LAFFERTY; MCCALLUM; PEREIRA, 2001). Jin e Ho (2009) aplicaram um modelo léxico baseado em HMM para identificar padrões e extrair aspectos e expressões de opinião. Jakob e Gurevych (2010) aplicaram CRF em *posts* de diversos domínios. Como atributos foram utilizados *tokens*, POS tags, dependência sintática, distância entre palavras, palavras de opinião. Yu et al. (2011) utilizaram modelos baseados em SVM para classificar os aspectos em um *post*. Cada instância representa um aspecto que é classificado como aspecto e não-aspecto. Todos os aspectos classificados são extraídos e representados através de sumários. O método também identifica expressões e sinônimos de aspectos frequentes. Entretanto, o método apresenta uma limitação importante: a identificação de aspectos só ocorre a partir de uma lista limitada de substantivos frequentes.

A *extração usando modelos de tópicos* utiliza modelos estatísticos para identificar tópicos em grandes *datasets*. A modelagem de tópicos é um método de aprendizagem de máquina não-supervisionado que considera que cada *post* consiste de uma mistura de

tópicos e cada tópico é uma distribuição probabilística de palavras. Cada *post* possui uma probabilidade de ter sido gerado por um determinado modelo de tópico. Existem duas estratégias básicas para a modelagem de tópicos: *Probabilistic Latent Semantic Analysis* (pLSA) e *Latent Dirichlet allocation* (LDA) Modelos de tópicos podem ser aplicados para a extração de aspectos, entretanto, tópicos podem conter aspectos e palavras de opinião e na análise de sentimentos estes precisam estar separados (LIU, 2012). Mukherjee e Liu (2012) utilizaram palavras-sementes definidas pelo usuário sobre algum aspecto para produzir distribuições de tópicos. Os autores também apresentam uma estratégia para separar palavras de sentimento e aspectos. A limitação deste trabalho, assim como em todos os métodos não-supervisionados, está na estimativa dos parâmetros que geralmente dependem de análise dos resultados através de diversas execuções.

Recentemente, estudos diretamente relacionados à análise de sentimentos baseada em aspectos são descritos por Pontiki et al. (2015). Toh e Su (2015) trata a extração de aspectos como um problema de classificação no qual modelos de classificação binários são treinados para prever a presença de aspectos. Como atributos, os autores utilizam BOW, POS *tags*, e modelagem de tópicos baseados em dados extraídos da Web (tais como Amazon¹ e Yelp²). Saias e Ramalho (2015) utilizaram o classificador MaxEnt (Máxima Entropia) para detectar entidades e aspectos. Como atributos foram utilizados lemas e padrões de pontuação. Rodrigues et al. (2016) propuseram um método para a classificação de polaridade em documentos financeiros. Os autores utilizam um estratégia de multi-modelos onde cada um descreve uma entidade.

Em relação ao trabalho desenvolvido, a extração de aspectos é realizada através de modelos de aprendizagem supervisionada. Estes modelos identificam a presença de menções explícitas ou implícitas a entidades ou aspectos de entidade. A descrição completa do método desenvolvido é apresentada na Seção 4.2.

3.2 Atribuição de Polaridade

A atribuição de polaridade é uma tarefa crucial na análise de sentimentos. Nesta Seção, serão apresentados os estudos fundamentais para esta tarefa bem como os trabalhos mais relevantes e recentes. Conforme mencionado no Capítulo 2, os *posts* podem expressar sentimentos em diferentes níveis. A análise de sentimentos em redes sociais

¹Disponível em: <http://www.cs.jhu.edu/~mdredze/datasets/sentiment/>

²Disponível em: http://www.yelp.com/dataset_challenge/

frequentemente faz análise em nível de documento, devido à complexidade envolvida na análise em nível aspectos e no grande volume de dados (LIU, 2012). Embora existam diversas formas de atribuir polaridade a *posts* (tais como, aprendizado supervisionado, aprendizado não-supervisionado, e dicionários), nesta seção, serão discutidos os estudos voltados à aprendizagem supervisionada, pois nesta categoria se enquadra a contribuição deste trabalho.

Redes sociais como o Twitter têm sido foco de diversos estudos por possuir duas características: (i) conteúdos curtos (até 140 caracteres); e (ii) textos que geralmente expressam opiniões (positivas ou negativas). Assim, usualmente busca-se classificar *tweets* como positivos ou negativos. Recentemente, diversas abordagens com esta finalidade foram propostas. Martínez-Cámara et al. (2014) apresentam uma estudo exaustivo sobre este assunto. A tarefa de identificar polaridade em *tweets* é tipicamente tratada como um problema de classificação. A classificação supervisionada requer *datasets* rotulados para treinamento e estes *datasets* geralmente exigem grande esforço manual para serem obtidos.

Go, Bhayani e Huang (2009) realizaram a classificação de sentimentos no Twitter com o objetivo de avaliar o sentimento de *tweets* como positivos e negativos. Os autores avaliaram a capacidade do uso de *emoticons* para determinar o sentimentos dos *tweets*, comparados a *tweets* anotados manualmente. Esta metodologia é interessante pois permite a criação de *datasets* rotulados automaticamente, porém adiciona bastante ruído aos dados. Os autores avaliaram diversos modelos de aprendizagem de máquina e obtiveram acurácia de 80% sobre os dados anotados automaticamente. Com o mesmo propósito, Barbosa e Feng (2010) utilizam sites detecção de sentimentos em *tweets*, tais como: Twendz, Twitter Sentiment e TweetFeel. Os autores utilizaram como atributos BOW, POS *tags* e informações sintáticas de *tweets* tais como: *retweet*; *hashtag*; *reply*; *link*. Eles também avaliaram a influência do tamanho do conjunto de treinamento nos resultados e constataram que a utilização conjunta de seus atributos manteve a taxa de erro constante conforme aumenta o número de instâncias de treinamento.

Speriosu et al. (2011) estudaram a influência que os seguidores exercem no sentimentos das pessoas. Os autores consideram que pessoas podem influenciar outras, compartilhando opiniões similares sobre os mesmos tópicos. Para isso, foram construídos grafos com usuários, *tweets*, e emoções nos nodos e suas conexões baseadas no grafo de seguidores, *hashtags* e *emoticons*. Os autores concluíram que o uso de grafos baseadas em *hashtags* e *emoticons* ajudam a melhorar os resultados enquanto que grafos de usuários

pioraram as métricas de avaliação.

O trabalho de Saif, He e Alani (2012) apresenta um estudo sobre análise de semântica de sentimentos em *tweets*. Os atributos semânticos consistem de conceitos como pessoa, produto, cidade que representam entidades como Steve Jobs, iPhone, e Londres extraídos de *tweets*. O objetivo dos autores é apresentar que conceitos estão mais relacionados a sentimentos e que estas relações podem ajudar a determinar o sentimento de um *tweet* de forma semântica. Os autores concluíram que a remoção de *stopwords* pode impactar negativamente os resultados e que o uso de conceitos semânticos, especialmente através de interpolação, contribui para melhores resultados.

Ghiassi, Skinner e Zimbra (2013) apresentam uma abordagem semi-automática para a redução de atributos em *tweets* utilizando *n*-gramas e análise estatística. Eles desenvolveram um dicionário com 187 atributos baseados em heurísticas que representam a cobertura de 97.3% do corpus a partir da simplificação de 755 regras. Os autores argumentam que estes atributos reduzem a complexidade do modelo de classificação e produzem melhores resultados. Os *tweets* utilizados tratam uma mesma entidade (Justin Bieber) e possuem classes desbalanceadas (82% dos dados são "positivos" enquanto 18% "negativos") isso pode explicar a revocação de cerca de 30% para negativos. Além disso, é difícil verificar a eficácia do método, pois os dados não são disponibilizados e o método não foi aplicado a dados comumente utilizados no estado da arte.

Coletta et al. (2014) exploraram a combinação de classificadores e *cluster ensembles* para a classificação de sentimentos de *tweets*. Semelhante à abordagem proposta, os autores consideram que *posts* semelhantes tendem a compartilhar o mesmo sentimento. No entanto, o método precisa estimar dois parâmetros: a importância relativa para os *clusters* e o número de iterações do algoritmo. A estimativa destes parâmetros é um problema de otimização. Diferente do método proposto, os autores utilizaram um dicionário contendo 6.789 palavras de sentimentos (HU; LIU, 2004) e frequência de *emoticons*. O uso de dicionários apresenta a desvantagem destes serem estáticos e a linguagem na Web é muito dinâmica, o que gera o problema conhecido como *out of vocabulary words*.

Carvalho, Prado e Plastino (2014) propõem uma estratégia estatística para a classificação das opiniões expressas em *tweets*, utilizando palavras de referência selecionadas por meio de um algoritmo genético. O método identifica um conjunto de possíveis palavras de referência a partir de palavras cujas polaridades são bem definidas (por exemplo, *good* e *bad*). Os resultados dos autores evidenciaram que palavras de referência selecionadas através de um critério estatístico melhoram a acurácia da classificação.

Weren et al. (2014) analisaram diversos atributos frequentemente utilizados na análise de perfil de autoria. Os autores concluíram que atributos extraídos de um ranking gerado a partir de um motor de busca são suficientes para classificar idade e gênero de autores, dado um texto. Similar a este trabalho, os autores utilizam funções de agregação simples para gerar atributos sobre o ranking gerado.

Silva, Hruschka e Jr. (2014) analisaram o uso de *feature hashing* para resolver o problema de esparsidade quando palavras são utilizadas como atributos. A *feature hashing* produz atributos mais discriminativos a partir de um vetor de *tokens* que compõem um *tweet*. Os autores reportam que sua técnica melhorou os resultados quando comparada à métodos baseados em frequência de *tokens*.

Saif et al. (2015) apresentam uma abordagem baseada em dicionários que identifica padrões de coocorrências de palavras de sentimentos. Estes padrões ajudam a descrever semanticamente a intensidade e a polaridade do sentimento atribuído a um *tweet*. Semelhante ao método desenvolvido, os autores consideram que palavras que ocorrem em contextos similares tendem ter significados similares. No entanto, os autores utilizam métodos complexos para o processamento de termos do índice, tais com POS *tags*, tratamentos de palavras estrangeiras e negação.

Os trabalhos relacionados utilizam palavras dos *posts* como atributos (BOW, geralmente unigramas e bigramas) (GO; BHAYANI; HUANG, 2009; DAVIDOV; TSUR; RAPPOPORT, 2010; SAIF; HE; ALANI, 2012), combinadas com dicionários (ZHANG et al., 2011; SPERIOSU et al., 2011). *Part-of-speech* (POS) *tags* também têm sido amplamente utilizadas (BARBOSA; FENG, 2010; GO; BHAYANI; HUANG, 2009; HU et al., 2013; SAIF; HE; ALANI, 2012). *Emoticons* e atributos específicos de *tweets* (tais como *retweets*, *hashtags* e grafo de seguidores) também foram explorados (SPERIOSU et al., 2011). Recentemente, *feature hashing* (SILVA; HRUSCHKA; JR., 2014) e *cluster ensembles* (COLETTA et al., 2014; SILVA; HRUSCHKA; JR., 2014; MARTÍNEZ-CÁMARA et al., 2014) também foram propostos.

Segundo Liu (2012), o ponto fundamental para a classificação de sentimentos está na seleção de atributos que sejam efetivamente significativos. A utilização de métodos baseados em palavras seja BOW ou dicionários possuem limitações, tais como palavras de sentimentos infrequentes e palavras fora do dicionário, respectivamente. Além disso, recursos linguísticos tais como POS *tags* geralmente dependem de textos escritos corretamente para serem efetivos, o que não é frequente em conteúdos da Web. O método desenvolvido visa eliminar essas dificuldades utilizando atributos que não empregam pa-

lavras de opinião como indicadores de sentimento. Ao invés disso, utilizamos apenas 24 atributos que visam identificar estruturas latentes entre o conteúdo e o sentimento atribuído.

3.3 Sumário dos Trabalhos Relacionados

A Tabela 3.1 apresenta os principais trabalhos relacionados e os tipos de atributos utilizados. Consideramos os seguintes critérios: *Bag-of-Words* (BOW) como sendo métodos que utilizam as palavras do texto como atributos; POS *tags* quando o método utiliza ferramentas de análise gramatical (etiquetador gramatical) como atributos ou para adicionar restrições; Dicionários quando os autores criam ou utilizam dicionários de sentimentos anotados manualmente para adicionar informações *a priori*; Relações de dependência quando utilizam análise sintática (árvores de dependência); o termo "outros" refere-se a atributos alternativos que não se enquadram nos tipos anteriores.

Analisando os trabalhos, de uma forma geral, o uso de BOW tem sido amplamente explorado. No entanto, estes atributos tendem ser representados através de vetores esparsos que resultam em modelos de classificação complexos. POS *tags*, da mesma forma está presente em grande parte dos trabalhos, principalmente em métodos voltados à identificação de aspectos. Dicionários têm sido amplamente utilizados para métodos que buscam atribuir a polaridade. Relações de dependência, por outro lado, tem mais afinidade com métodos para identificação de aspectos. Finalmente, métodos alternativos são mais frequentes em trabalhos mais recentes.

Os métodos para identificação de aspectos dependem de vários recursos linguísticos e a nossa proposta nessa área não foge a esta regra. Contudo, para a atribuição de polaridade, conseguimos uma alternativa de baixo custo que não depende de recursos linguísticos. Na Tabela 3.1, ASPECTOS refere-se ao método desenvolvido para identificação de aspectos; e POLARIDADE refere-se ao método que atribui a polaridade a um *post*.

Tabela 3.1: Tipos de atributos utilizados no estado da arte

Trabalho	Bag-of-Words	POS tags	Dicionários	Relações de dependência	Outros
Hu e Liu (2004)	X	X		X	
Moghaddam e Ester (2010)	X	X			
Qiu et al. (2011)	X	X		X	
Liu et al. (2013)	X	X	X	X	
Jin e Ho (2009)	X	X	X		
Jakob e Gurevych (2010)	X	X	X	X	
Yu et al. (2011)	X	X		X	
Mukherjee e Liu (2012)	X	X	X	X	X
Toh e Su (2015)	X	X	X	X	
Saias e Ramalho (2015)	X	X	X	X	X
Rodrigues et al. (2016)	X	X	X	X	X
Martínez-Cámara et al. (2014)		X	X		X
Go, Bhayani e Huang (2009)	X	X			X
Barbosa e Feng (2010)		X	X		X
Speriosu et al. (2011)			X		X
Saif, He e Alani (2012)	X	X	X		X
Coletta et al. (2014)	X				X
Carvalho, Prado e Plastino (2014)	X				X
Ghiassi, Skinner e Zimbra (2013)	X		X		X
Silva, Hruschka e Jr. (2014)	X		X		X
Saif et al. (2015)	X	X	X		
ASPECTOS	X	X	X	X	X
POLARIDADE	*		*		X

Fonte: O autor

4 EXTRAINDO ENTIDADES E ASPECTOS ATRAVÉS DA CLASSIFICAÇÃO EM DUAS FASES

Neste capítulo será descrito o método desenvolvido para a identificação e extração de categorias de opiniões em *posts*. O método desenvolvido combina algoritmos de classificação, ferramentas de resolução de correferências, analisador gramatical e outros recursos definidos na Seção 2.3. O objetivo deste método é minimizar o uso de recursos externos, tais como dicionários, priorizando informações explícitas no *post* para a identificação e extração de categorias e alvos da opinião.

Os experimentos descritos neste capítulo foram realizados no âmbito da campanha de avaliação *Aspect-based sentiment analysis* do SemEval-2015 (PONTIKI et al., 2015). O objetivo da campanha é promover a análise de sentimentos baseada em aspectos. Na edição de 2015 foram apresentadas as tarefas de extração de categorias e alvos da opinião e classificação de polaridade. O evento atraiu 16 equipes de pesquisa que enviaram seus métodos. O método apresentado neste capítulo foi submetido a esta avaliação (KAUER; MOREIRA, 2015), onde obteve pequena vantagem sobre o *baseline* em todas as tarefas. O método proposto foi classificado em 7 de 10 para identificação de categorias; 9 de 13 para a extração do alvo da opinião; 5 de 7 para a identificação conjunta da categoria e alvo; e 5 de 14 para polaridade.

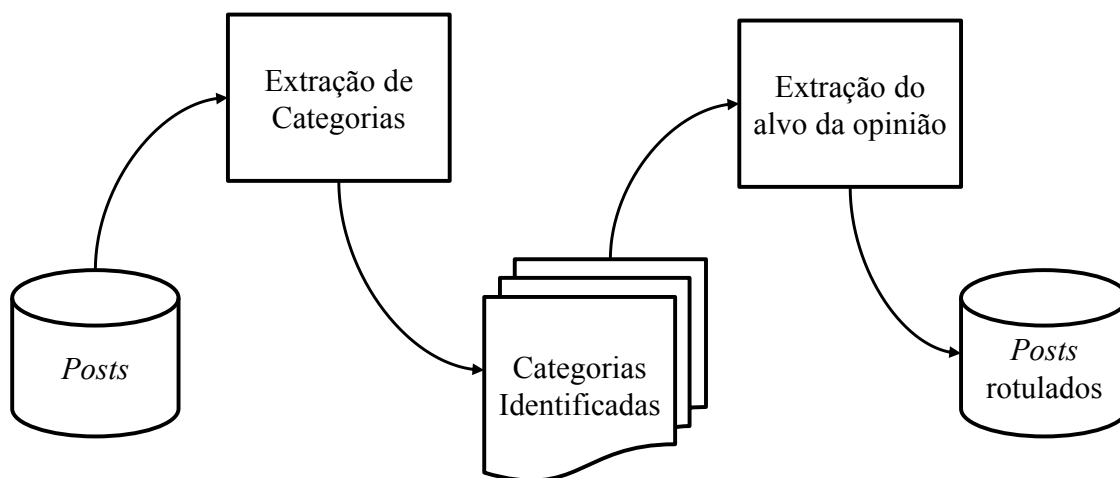
4.1 Visão Geral do Processo

A abordagem desenvolvida é composta por dois passos principais. O primeiro passo consiste em treinar um modelo de classificação para cada uma das categorias pertencentes a um domínio. No segundo passo, as palavras candidatas de uma categoria são utilizadas para identificar o alvo da opinião. A atribuição de polaridade será detalhada no Capítulo 5. A Figura 4.1 apresenta a visão geral do processo utilizada no método desenvolvido.

4.2 Categorização de opiniões

Conforme definido no Capítulo 2, uma categoria contém duas partes: a entidade e o aspecto da entidade. A estratégia de identificação de categorias consiste em duas etapas:

Figura 4.1: Visão Geral do Processo



Fonte: O autor

primeira, identificação da entidade; e segunda, a identificação do aspecto da entidade. Além disso, considera-se que um *post* contém uma ou mais sentenças e as sentenças contém zero ou mais opiniões.

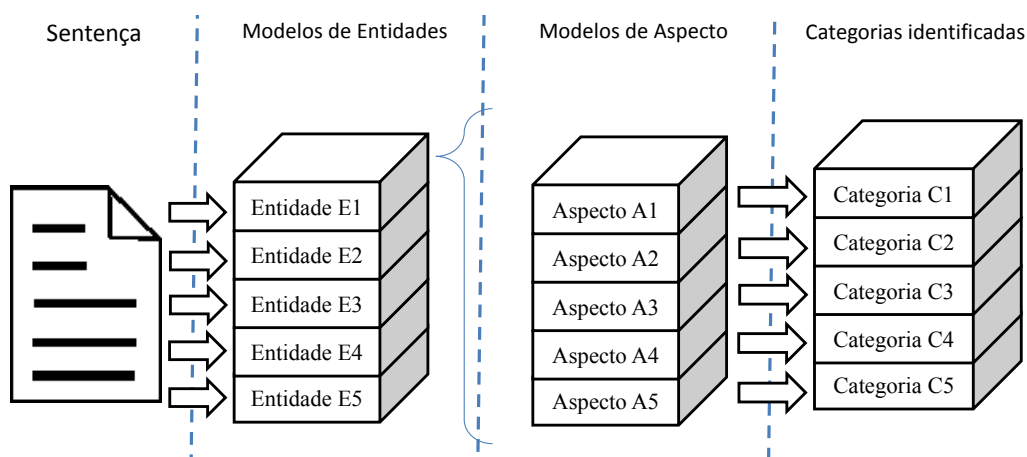
Para a identificação de categorias, é treinado um modelo de classificação para cada uma das categorias possíveis no domínio. Estes modelos de classificação indicam se uma determinada sentença contém opiniões de uma determinada categoria. Visto que é possível que uma sentença contenha mais de uma categoria, todas as categorias devem ser verificadas para cada sentença.

O fluxo de identificação de categorias consiste em duas partes: a classificação da presença da entidade, utilizando os modelos de entidade; e a classificação de aspectos, através dos modelos de aspectos. Se uma sentença não contém opiniões de uma entidade, não há necessidade de verificar a presença de aspectos da entidade, como forma de otimizar o processo. A Figura 4.2 apresenta este fluxo da identificação de categorias.

Cada sentença do *dataset* é representada através de instância no modelo de classificação. A transformação de um *post* para um conjunto de instâncias é apresentada na Figura 4.3 e envolve 6 etapas principais: Extração de lemas; Identificação de correferências; Representação vetorial; *InformationGain*; Seleção de atributos; e Transformação da primeira fase.

Extração de lemas. A partir de um *post* original (Figura 4.3a), todos os *tokens* são substituídos por seus respectivos lemas (Figura 4.3b). Nesta etapa, os lemas são extraídos seguindo a metodologia implementada por Manning et al. (2014).

Figura 4.2: Fluxo para a categorização de opiniões



Fonte: O autor

Identificação de correferências. Identifica expressões no *post* que se referem a mesma entidade ou aspecto. Ao utilizar correferências, são adicionadas palavras mais significativas a aspectos que podem estar em outras sentenças, aumentando a revocação. Na Figura 4.3c, a palavra *iPhone* é adicionado à sentença S_2 , por ser mencionada através da palavra *It*.

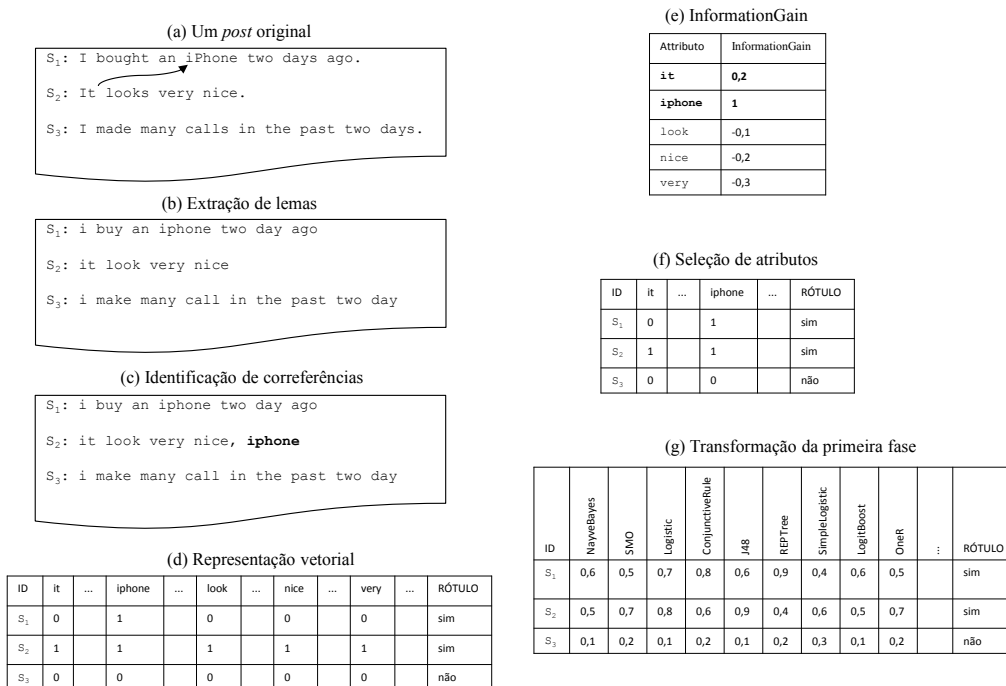
Representação vetorial. Cada sentença é convertida em uma instância. Uma instância é um vetor, onde cada posição é um atributo que contém a frequência do *token* na sentença. A última posição, chamada de rótulo ou classe, indica se a sentença contém determinada entidade ou aspecto. Um exemplo desta representação é mostrado na Figura 4.3d.

InformationGain. Nesta etapa, é utilizado um método de seleção de atributos chamado *InformationGain* (descrito na Seção 2.3.2). Nesta etapa é calculada a importância de cada atributo para a determinação da classe (Figura 4.3e).

Seleção de atributos. Nesta etapa, os atributos irrelevantes são descartados através da definição de um limiar mínimo de importância. Em nosso método, utilizou-se como limiar ($InfoGain \geq 0$), que considera somente atributos (*tokens*) positivamente correlacionados com a classe a ser prevista. Após esta etapa somente os atributos mais discriminativos são preservados, reduzindo a quantidade de atributos (Figura 4.3f).

Transformação da primeira fase. Utiliza a saída de classificação de uma sentença como atributos para a classificação. Este processo será detalhado no decorrer das próximas Subseções. Para este trabalho, foram desenvolvidas duas estratégias para a identificação de entidades e aspectos da entidade, chamados de classificação em uma fase e classificação em duas fases (Figura 4.3g).

Figura 4.3: Etapas de transformação de sentenças em instâncias



Fonte: O autor

4.2.1 Classificação em uma fase

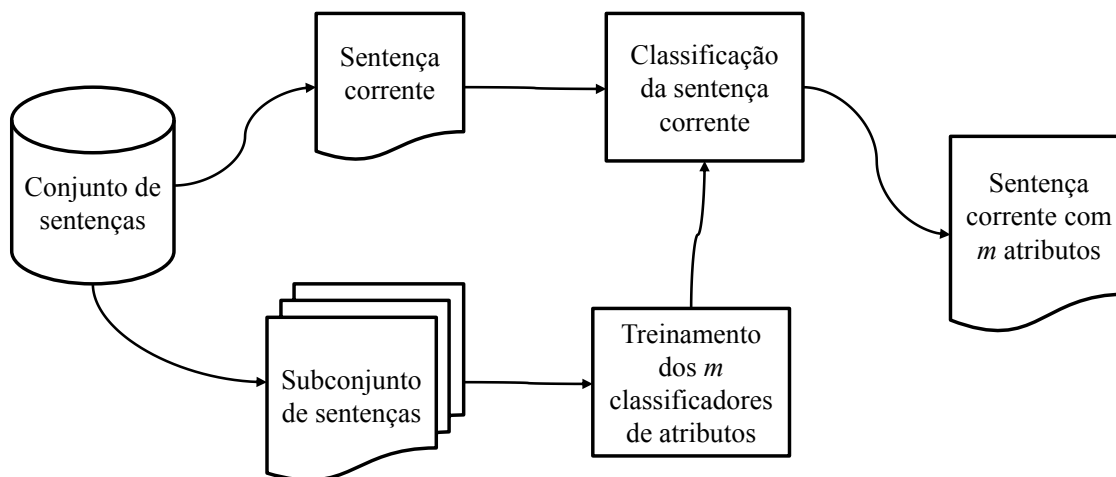
Para esta classificação são treinados l modelos para cada entidade usando todos as instâncias do domínio. Estes também serão chamados de *Modelos de Categoria* pois são responsáveis por determinar se uma sentença contém uma entidade ou aspecto da entidade. Para cada aspecto de uma entidade também é criado um conjunto de l modelos de classificação. Entretanto, o modelo é restrito somente a sentenças que contêm opiniões de mesma entidade.

4.2.2 Classificação em duas fases

Para esta classificação foram treinados m modelos para cada entidade e aspecto, utilizando todas as sentenças exceto uma (chamada sentença corrente). Estes modelos são chamados de *Modelos de Atributo*. Na primeira fase, para cada sentença (corrente) são gerados m modelos de classificação e a sentença corrente é classificada. Na segunda fase, são selecionadas as saídas de classificação de cada um dos m modelos para serem utilizados como atributos para os Modelos de Categoria.

A primeira fase exige grande esforço de processamento, visto que um novo conjunto de treinamento é gerado para cada sentença e o modelo de classificação é atualizado. Este método também reduz a dimensionalidade do modelo, convertendo os atributos específicos de uma sentença para informações genéricas da sentença em relação ao domínio dos dados. O fluxo deste processo é apresentado na Figura 4.4.

Figura 4.4: Processo de geração de atributos em duas fases



Fonte: Adaptado de Kauer e Moreira (2015)

Para realizar a classificação de uma sentença do conjunto de testes, a sentença é processada e então os seus lemas e correferências são identificados. Após, as frequências de cada lema é selecionada e os m Modelos de Atributos produzem os valores para os atributos para a segunda fase. A classe final, que informa se a categoria avaliada foi mencionada na sentença, é definida pelo maior escore obtido na soma da saída dos l Modelos de Categoria. Em nossos experimentos os valores de m e l são definidos como 20 e 6, respectivamente.

4.3 Identificação do alvo da opinião

Embora uma sentença possa conter uma ou mais categorias, também pode conter zero ou mais alvos da opinião que são palavras ou expressões que caracterizam uma entidade ou aspecto da entidade. Embora muitas abordagens adotem a identificação do alvo da opinião antes da Categorização, em nosso método o alvo é identificado após a categoria ser definida. Nós escolhemos esta abordagem principalmente por utilizarmos métodos baseados em *bag-of-words* e pelo fato de que algumas categorias não dispõem de alvos

explícitos no texto mesmo utilizando resolução de correferências.

Assim, para cada categoria identificada, o método seleciona as palavras candidatas da sentença de forma ordenada por sua importância (*InfoGain*) para a definição do aspecto concatenadas com as palavras do modelo da entidade. Esta abordagem supõe que as palavras oriundas do aspecto são mais significativas do que as palavras da entidade (que tendem a ser mais genéricas).

Em muitos casos, palavras pouco frequentes não são selecionadas como candidatas. Portanto, também são selecionadas palavras diretamente associadas às palavras candidatas através de uma relação gramatical. São consideradas relações gramaticais relevantes nesta etapa as relações do tipo *modificador adjetivo* (amod), *substantivo composto* (nn), e *sujeito nominal* (nsubj). A lista completa de relações gramaticais está descrita na Tabela 2.2.

O alvo da opinião é geralmente representado por um substantivo, conforme amplamente adotado na literatura (HU; LIU, 2004; QIU et al., 2011; LIU, 2012; LIU et al., 2013; PONTIKI et al., 2015). Assim, as potenciais classes gramaticais para alvos da opinião são NN (substantivo singular) e NNS (substantivo plural). Para detectar erros de identificação é utilizada uma lista de 5000 palavras foi produzida por Liu et al. (2013). Esta lista de exceções contém palavras com pouco ou nenhum significado específico que tipicamente não representam alvos válidos. Candidatos a alvo da opinião que constam dessa lista são descartados, desta forma, o alvo principal é o primeiro candidato que não estiver nessa lista.

Se nenhum substantivo for encontrado nesses candidatos, são selecionados os substantivos que estão indiretamente relacionados, através das relações gramaticais, aos candidatos. Se após esta verificação nenhum substantivo for identificado, por exemplo, nenhum substantivo válido para uma determinada categoria, o alvo é considerado "NULL". Alguns alvos são compostos por expressões, estas expressões são obtidas usando relações gramaticais do tipo *substantivo composto* (nn).

Uma limitação deste método é que não são identificados múltiplos alvos na mesma sentença para uma mesma categoria. Portanto, considera-se que cada categoria encontrada tem somente um alvo, o que pode não ser verdade na prática, por exemplo, quando um mesmo alvo é referenciado de formas diferentes na mesma sentença. Entretanto, visto que uma sentença pode conter diversas categorias, diversos alvos podem ser encontrados, inclusive podendo ser um mesmo em categorias distintas.

4.4 Avaliação Experimental

Os experimentos relatados a seguir incluem os métodos descritos anteriormente neste Capítulo. Inicialmente são descritas as configurações dos experimentos, tais como: pré-processamento, ferramentas, e *datasets* utilizados. Após, são descritos os *baselines* e a metodologia de avaliação. Finalmente, são apresentados os resultados obtidos com o método.

4.4.1 Configuração dos Experimentos

Conteúdos gerados por usuários caracterizam-se pela presença ruído no texto, seja pelo uso excessivo de pontuação, gírias ou erros de ortografia. Esta característica faz com que a grande parte das ferramentas de processamento de linguagem natural esteja suscetível a erros que impactam diretamente no resultado final do método. No método desenvolvido, foi utilizada a StanfordNLP Core Toolkit (SCT)(MANNING et al., 2014)¹ que oferece um conjunto de ferramentas necessário para todas as etapas do pré-processamento, incluindo:

- `Tokenize` realiza o particionamento do *post* em *tokens* (geralmente palavras);
- `Sentence Split` identifica as sentenças no *post*;
- `Part of Speech` é um analisador sintático, identifica a classe gramatical de cada *token*;
- `Lemma` identifica a formação básica de cada *token* (lema);
- `Dependency Parsing` identifica a árvore de dependência entre *tokens* da mesma sentença;
- `Sentiment Analysis` identifica o sentimento atribuído a um *token*;
- `Coreference` identifica as correferências na entre *tokens* do mesmo *post*.

Esta ferramenta realiza todas as etapas de pré-processamento, reduzindo assim o número de ferramentas necessárias para a reprodução do método. Embora o conjunto de treinamento utilizado esteja segmentado por sentenças manualmente, muitas vezes ocorrem caracteres de pontuação no interior das sentenças. Assim, antes de aplicar o StanfordNLP, foi realizada uma substituição de caracteres de pontuação (.!?) por vírgula,

¹Disponível em: <http://stanfordnlp.github.io/CoreNLP/>

tornando a sentença compatível com a ferramenta. Uma causa frequente de erros de processamento é ocasionada pela presença de números e caracteres não-alfabéticos, por isso estes também foram removidos.

Os experimentos foram executados sobre três domínios disponíveis em Pontiki et al. (2015), chamados de Restaurantes (Res), Laptops (Lap) e Oculto (Hid), domínio desconhecido. A Tabela 4.1, apresenta um resumo dos *datasets* utilizados.

Tabela 4.1: Detalhes dos datasets

<i>Dataset</i>	<i>#posts</i>	<i>#Sentenças</i>	<i>#Opiniões</i>	<i>#Alvos</i>	<i>#Ent.</i>	<i>#Asp.</i>	<i>#Cat.</i>
Res Train	254	1315	1654	527	6	5	13
Res Test	96	685	845	270	6	5	12
Lap Train	277	1739	1974	0	22	9	81
Lap Test	173	761	949	0	19	9	58
Hid Test	30	266	339	92	7	8	28

Fonte: Pontiki et al. (2015)

4.4.2 Metodologia de avaliação e *baselines*

Para a avaliação do método seguiu-se a metodologia proposta em Pontiki et al. (2015), onde utilizou-se o conjunto de métricas definidas na Seção 2.4: precisão (Pr), revocação (Rec) e Medida-F (F1). Assim, a avaliação compreende três etapas descritas a seguir:

1. **Identificação de categorias:** A avaliação verifica se o método identifica e retorna um conjunto de categorias expressas pela opinião. Para isso são calculadas precisão, revocação e Medida-F pela comparação das lista de categorias obtidas pelo método (por sentença) que correspondem a lista esperada. O cálculo ignora ocorrências de categorias duplicadas. Categorias não identificadas são consideradas como *FN*, enquanto que categorias identificadas fora do conjunto esperado são considerados como *FP*. Esta etapa foi aplicada para os domínios de Restaurantes e Laptops.
2. **Identificação do alvo:** A avaliação verifica se o método identifica e retorna um conjunto de alvos, ou seja, as expressões que são usadas na sentença para se referir a uma entidade ou aspecto da entidade. Para isso são calculadas precisão, revocação e Medida-F pela comparação das lista de alvos obtidos pelo método (por sentença) que correspondem a lista esperada. Esta lista é construída pelos valores da posição

absoluta no texto. O cálculo ignora ocorrências "NULL", desde que não correspondam a menções explícitas do alvo. Assim como na classificação de categorias, alvos não identificados são considerados como FN , enquanto que alvos identificados fora do conjunto esperado são considerados como FP . Esta avaliação foi aplicada somente para o domínio de Restaurantes.

3. **Identificação combinada do par categoria e alvo:** A avaliação considera que se o método identificou uma categoria e um alvo. Novamente, são calculadas precisão, revocação e Medida-F pela comparação dos pares obtidos pelo método (por sentença) que correspondem a lista esperada. Esta avaliação foi aplicada somente para o domínio de Restaurantes.

Os *baselines*, da mesma forma, foram implementados por Pontiki et al. (2015). Eles são descritos a seguir:

1. **Identificação de Categorias:** O *baseline* consiste de um modelo de *Support Vector Machine* (SVM) com *kernel* linear. Como atributos foram considerados z *bag-of-words* (BOW) mais frequentes de cada sentença que contém determinada categoria. Para cada sentença de teste é verificada a probabilidade de pertencer a uma determinada categoria.
2. **Identificação do Alvo:** O *baseline* usa o conjunto de treinamento usado para cada categoria. São extraídos todos os alvos possíveis desta categoria. Para uma sentença de teste pertencente a uma determinada categoria, o alvo é o primeiro pertencente a lista gerada para aquela categoria. Se não existir na lista, é assinado o valor "NULL".

Para o número de z de termos frequentes foi definido em 1000 para cada SVM treinado. As pontuações dos *baselines* foram calculadas sobre a décima parte para testes e as restantes nove partes para treinamento (proporção 90/10). Os resultados dos *baselines* sobre o conjunto de treinamento são apresentados na Tabela 4.2.

Tabela 4.2: *Baselines* para Identificação de Categorias e Alvos

Domínio	Informação Avaliada	Precisão	Revocação	Medida-F
Restaurantes	Categoria	0,53	0,68	0,60
Restaurantes	Alvo	0,55	0,43	0,48
Restaurantes	(Categoria, Alvo)	0,36	0,42	0,38
Laptops	Categoria	0,35	0,41	0,38

Fonte: Pontiki et al. (2015)

4.4.3 Identificação de Categorias e Alvos da Opinião

Para esta avaliação foram consideradas precisão, revocação e Medida-F. A Tabela 4.3 apresenta os resultados obtidos usando nossa abordagem comparado ao *baseline* para identificação de categorias. A Tabela 4.4 mostra os resultados obtidos para a identificação do Alvo da opinião. Os resultados para o par (categoria, alvo) são apresentados na Tabela 4.5.

Tabela 4.3: Identificação de Categorias

Domínio	Método	Precisão	Revocação	Medida-F
Restaurantes	2Fases	0,6556	0,4323	0,5210
Restaurantes	1Fase	0,6835	0,4181	0,5188
Restaurantes	1Fase- <i>coref</i>	0,6821	0,4180	0,5184
Restaurantes	<i>baseline</i>			0,5133
Restaurantes	2Fases- <i>coref</i>	0,6509	0,4090	0,5023
Laptops	1Fase- <i>coref</i>	0,4834	0,4462	0,4640
Laptops	<i>baseline</i>			0,4631
Laptops	2Fases- <i>coref</i>	0,4689	0,4388	0,4534
Laptops	1Fase	0,5066	0,4040	0,4495
Laptops	2Fases	0,4773	0,4209	0,4473

Fonte: Kauer e Moreira (2015)

O método desenvolvido tiveram pequeno ganho sobre o *baseline* em ambas as abordagens para o domínio de Restaurantes. Para este domínio, a abordagem de duas fases (2Fases) foi superior a abordagem de uma fase (1Fase). Para o domínio de Laptops, no entanto, o método desenvolvido obteve resultados inferiores ao *baseline*. Esta redução nos resultados pode ser atribuída à dificuldade de identificar correferências nestas sentenças, principalmente devido ao grande número de palavras fora do vocabulário (CPU, HD, RAN, etc). Além disso, devido à grande quantidade de categorias possíveis combinada ao pequeno número de instâncias de treinamento dificulta a generalização do modelo. A Tabela 4.3 também mostra melhores resultados quando não é aplicada a resolução de correferências (*-coref*) no domínio de Laptops, embora para o domínio de Restaurantes o uso de correferências tenha contribuído para melhores resultados.

Considerando os resultados para detecção do alvo da opinião, ambas versões do método superaram o *baseline*. A classificação em duas fases obteve melhor revocação na identificação de categorias e alvos da opinião, mas piores resultados em termos de precisão comparados a classificação em uma fase.

Tabela 4.4: Identificação do Alvo da Opinião

Domínio	Método	Precisão	Revocação	Medida-F
Restaurantes	2Fases	0,5656	0,4373	0,4932
Restaurantes	2Fases- <i>exc</i>	0,5632	0,4354	0,4911
Restaurantes	1Fase	0,5764	0,4244	0,4888
Restaurantes	1Fase- <i>exc</i>	0,5739	0,4225	0,4867
Restaurantes	<i>baseline</i>			0,4807

Fonte: Kauer e Moreira (2015)

Foram executados alguns experimentos para avaliar a contribuição do uso da lista de exceções durante a identificação do alvo. Estas execuções estão identificadas como 1Fase-*exc* e 2Fases-*exc* na Tabela 4.4. Nestes experimentos constatou-se que seu uso que a lista não impacta nos resultados.

Tabela 4.5: Identificação do par (categoria, alvo)

Domínio	Método	Precisão	Revocação	Medida-F
Restaurantes	2Fases	0,4852	0,2722	0,3487
Restaurantes	<i>baseline</i>			0,3444
Restaurantes	1Fase	0,4521	0,2734	0,3407
Restaurantes	1Fase- <i>coref</i>	0,4694	0,2639	0,3378
Restaurantes	2Fases- <i>coref</i>	0,4496	0,2591	0,3288

Fonte: Kauer e Moreira (2015)

Para os resultados considerando o par (categoria, alvo) a classificação em duas fases ficou próximo ao *baseline* e a classificação em uma fase. O ganho em termos de precisão é de três pontos percentuais, com pequena redução em termos de revocação. A melhor configuração do método desenvolvido, portanto, foi utilizando resolução de correferências e utilização da lista de exceções.

4.4.4 Análise de Erros

O método desenvolvido superou o *baseline* com pequena vantagem na maioria dos casos, entretanto, para alguns casos o método apresentou resultados inferiores. A seguir serão listados alguns casos onde as escolhas do método contribuíram para o aumento no número de erros de classificação.

Para a tarefa de classificação de categorias, a escolha do parâmetro de limiar para *InfoGain* (≥ 0) pode ter impactado negativamente nos resultados. Entretanto, algum

método para seleção e ordenação de palavras candidatas é requerido para a abordagem desenvolvida. Além disso, a utilização de todas as palavras pode impactar diretamente no tempo de processamento.

Foram utilizadas as palavras selecionadas pelo *InfoGain* como candidatas para identificar o alvo da opinião. Nos experimentos realizados, em muitos casos, o alvo estava próximo às palavras selecionadas por esta estratégia. Isto acontece devido à classe *positivo* (para determinação de uma categoria) conter menos instâncias que a classe *negativo*, assim o *InfoGain* tende a selecionar palavras que caracterizem a classe menos frequente. No entanto, muitos erros de classificação aconteceram porque a estratégia falhou em identificar palavras infrequentes que correspondem à categoria esperada. Uma possível alternativa para mitigar este problema pode ser o uso de sinônimos.

4.4.5 Resultados Oficiais SemEval-2015

Os experimentos descritos neste Capítulo foram realizados no âmbito da campanha de avaliação SemEval-2015 (PONTIKI et al., 2015). Para fins de comparação, a Tabela 4.6 apresenta os resultados oficiais do evento para a identificação de entidades e aspectos. * refere-se a métodos que utilizam recursos externos ao conjunto de treinamento, tais como dicionários. UFRGS refere-se a nossa participação do evento. UFRGS* refere-se a nossa abordagem utilizando a classificação em duas fases.

Tabela 4.6: Resultados oficiais do SemEval-2015

Laptops		Restaurantes	
Equipe	F1	Equipe	F1
NLANGP	50,86*	NLANGP	62,68*
Sentiue	50,00*	NLANGP	61,94
IHS-RD.	49,59	UMDuluthC	57,19
NLANGP	49,06	UMDuluthT	57,19
TJUdeM	46,49	SIEL	57,14*
UFRGS	44,95	Sentiue	54,10*
UFRGS	44,73*	LT3	53,67*
V3	24,94*	TJUdeM	52,44*
		UFRGS	52,09*
		UFRGS	51,88
		IHS-RD.	49,87
		IHS-RD.	49,16
		V3	41,85*
Baseline	48,06	Baseline	51,32

Fonte: Pontiki et al. (2015)

5 ATRIBUIÇÃO DE POLARIDADE BASEADA EM RECUPERAÇÃO DE INFORMAÇÕES

Como mencionado no Capítulo 2, um *post* expressa a opinião do seu autor sobre uma ou mais entidades. Conforme discutido no Capítulo 4, métodos para a identificação de entidades e aspectos da entidade são bastante complexos e muitas vezes ineficientes para uma aplicação real. No entanto, após identificar as opiniões em um *post* é necessário atribuir a polaridade para o sentimento expresso.

O Twitter¹ é uma das redes sociais mais populares atualmente, onde os usuários podem expressar suas opiniões através de *posts* curtos (até 140 caracteres), chamados *tweets*. Estes *tweets* são interessantes para o estudo da análise de sentimentos, principalmente por tratar uma gama praticamente ilimitada de tópicos. A limitação de caracteres faz com que os usuários sejam mais objetivos em suas opiniões. Assim, tipicamente a análise de sentimentos em *tweets* é feita em nível de documento. Neste trabalho, consideramos que um *tweet* é um *post* que se refere à uma entidade e expressa um sentimento positivo, negativo ou neutro.

O método desenvolvido neste trabalho utilizou-se a seguinte hipótese: *posts* similares tendem a compartilhar a mesma polaridade do sentimento. Assim, informações sobre a polaridade dos n *posts* mais similares podem ajudar a classificar a polaridade para um *post* não rotulado. O método desenvolvido é semelhante ao apresentado por Weren et al. (2014) que aplica este conceito para classificar gênero e idade de textos de mídia social.

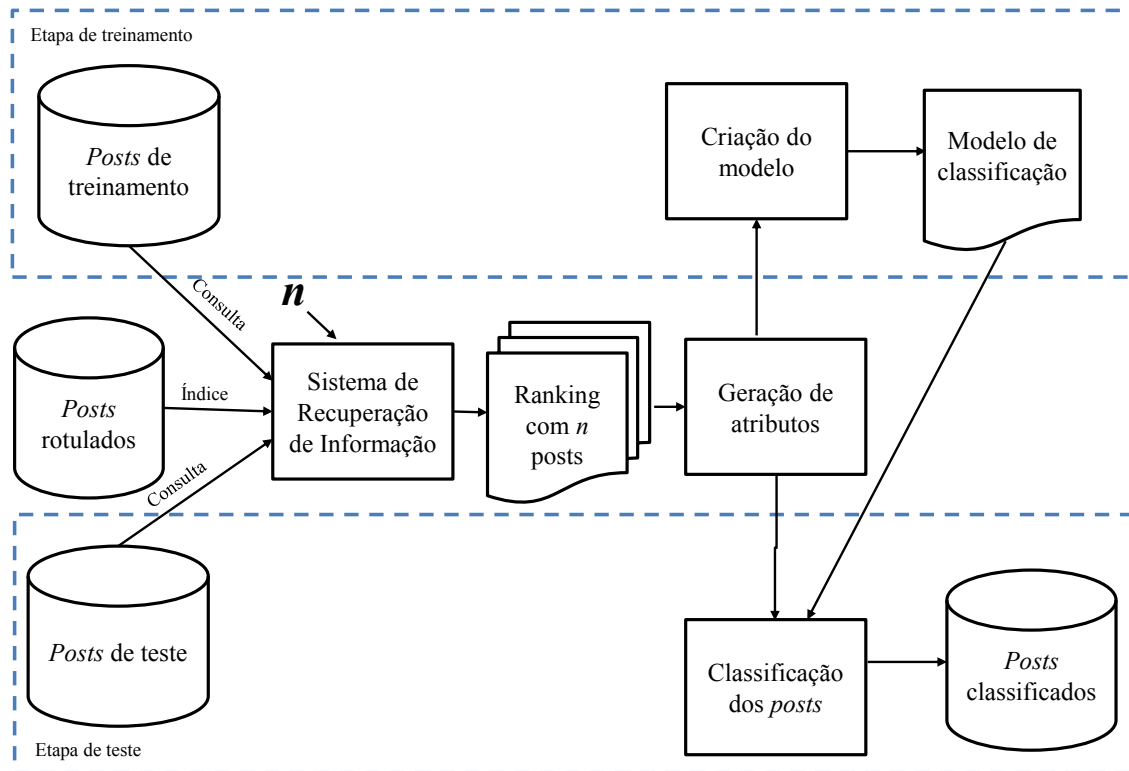
5.1 Visão Geral do Processo

O método desenvolvido é composta por dois passos. O primeiro passo consiste em obter os n *posts* mais similares ao *post* que se quer classificar. O segundo passo, utiliza a informação sobre estes n *posts* como atributos para treinar um modelo de classificação supervisionado. Uma visão geral do processo utilizado é apresentada na Figura 5.1. Formalmente, o método consiste em: dado um conjunto de *posts* $T = \{t_1, t_2 \dots, t_m\}$ para os quais a classe $c_i \in \{positive(+), negative(-)\}$ é conhecida e um conjunto de *posts* não rotulados $Q = \{q_1, q_2 \dots, q_p\}$, onde a classe é desconhecida. Foi utilizada a informação

¹Disponível em: <https://twitter.com>

sobre a similaridade de cada elemento $q_i \in Q$ em relação aos elementos $t_j \in T$ para classificar q_i . Informações quanto à similaridade entre *posts* são obtidas através de um Sistema de Recuperação de Informação.

Figura 5.1: Visão Geral do Processo



Fonte: O autor

O método pode utilizar um ou mais conjuntos de *posts* rotulados que são necessários para a criação do índice. O custo de indexação representa uma fração do custo de treinamento, assim este método pode ser aplicada para grandes conjuntos de dados. Nos experimentos realizados, foram utilizados três conjunto de dados: um grande conjunto de *posts* rotulados com ruído para indexação (*posts* rotulados); um conjunto manualmente anotado para treinamento (*posts* de treinamento); e um conjunto para teste (*posts* de teste).

Etapa de treinamento. O método utiliza um conjunto de testes, onde cada *post* é submetido ao Sistema de Recuperação de Informação em forma de consulta ao índice. A partir do *ranking* gerado, a Geração de atributos calcula os atributos para cada instância de treinamento que representa o *post*. Após gerar os atributos para as instâncias de treinamento, um modelo de aprendizagem supervisionada é treinado.

Etapa de teste. Assim como na Etapa de treinamento, cada *post* do conjunto de teste é submetido ao Sistema de Recuperação de Informação em forma de consulta ao mesmo

índice utilizado no treinamento. Os atributos são gerados a partir do *ranking* da mesma forma. A Classificação dos *posts* utiliza o Modelo de classificação gerado na Etapa de treinamento para prever o rótulo de cada *post* do conjunto de teste.

5.2 Sistemas de Recuperação de Informação

Um Sistema de Recuperação de Informação (SRI) utiliza funções de similaridade para ordenar um conjunto de itens (geralmente documentos textuais) como resultado de uma consulta, formando um *ranking*. Funções de similaridade calculam a pontuação de similaridade, ou score, entre dois itens. Neste método, um SRI é utilizado para indexar os *posts* rotulados em T e ordenar a relação para cada *post* não rotulado em Q . A lista ordenada (*ranking*) com n itens com maior similaridade para cada q_i serve para fonte de atributos que serão utilizados pelo modelo de classificação. Se um *post* q_i está próximo a *posts* positivos então ele pode ser positivo também. A ideia é semelhante ao princípio utilizado pelo algoritmo de classificação k-NN, no qual uma instância de teste é classificada como pertencente a maior classe dos k vizinhos mais próximos. Uma abordagem simples utilizaria este princípio, no entanto esta evidência isolada pode não ser suficiente. Assim, o método desenvolvido extrai diversas fontes de evidência do *ranking* composto pelos n *posts* mais similares a q .

O primeiro passo é a indexação dos *posts* rotulados (T) usando um SRI. Então, cada *post* não rotulado $q_i \in Q$ é utilizado como consulta e os n *posts* mais similares são recuperados e ordenados de forma decrescente por sua similaridade. A Tabela 5.1 apresenta um exemplo de *ranking* com $n = 25$.

5.3 Geração de Atributos

O modelo de classificação utiliza um conjunto de 24 atributos, sendo 12 para cada classe de sentimento. A partir do *ranking* gerado, funções de agregação como média, máximo, mínimo, soma e contagem são utilizadas para cada classe (positivo e negativo). A Tabela 5.2 apresenta os valores extraídos para o *ranking* da Tabela 5.1. Neste exemplo existem 12 itens negativos, então $count_- = 12$. A soma dos scores negativos é 260,68, representado pelo atributo sum_- . A média de scores para *posts* negativos é 21,723 e seu valor é representado pelo atributo avg_- . Adicionalmente, estas funções são combi-

nadas formando outros atributos, denominados ϕ , que consideram informações absolutas e relativas (em relação à classe) no *ranking* de cada item. O atributo ϕ ajuda a calcular a proximidade dos *posts* (positivos ou negativos) do topo do *ranking*. O cálculo de ϕ é definido como:

$$\phi_c = \sum_{r=1}^{n_c} \frac{rank_{rel}}{rank_{abs}} \quad (5.1)$$

onde c é a classe do sentimento; n_c é o número de *posts* recuperados para a classe, $rank_{rel}$ é a posição relativa para um *post* pertencente aquela classe, e $rank_{abs}$ é a posição absoluta do *post* no *ranking*.

O restante dos atributos são combinações da métrica ϕ e funções de agregação. São eles: $\phi_{avg_c} = \frac{\phi_c}{avg_c}$, $\phi_{max_c} = \frac{\phi_c}{max_c}$, $\phi_{min_c} = \frac{\phi_c}{min_c}$, $\phi_{sum_c} = \frac{\phi_c}{sum_c}$, e $\phi_{count_c} = \frac{\phi_c}{count_c}$.

As funções de agregação avg_c , max_c , min_c , e sum_c calculam os escores médio, máximo, e mínimo para a classe c , respectivamente; e $count_c$ é o número de *posts* para classe c no *ranking*. O último atributo, $\phi_{c_{positional}}$, combina os escores dos *posts* recuperados e sua posição no *ranking*. Ele é definido como:

$$\phi_{positional_c} = \sum_{r=1}^n \left(\frac{rank_{rel}}{rank_{abs}} \right) \times S(q_i) \quad (5.2)$$

onde $S(q_i)$ é o score para o *post* q_i . A Tabela 5.3 apresenta a descrição completa de todos os atributos gerados.

Uma particularidade importante da criação de atributos usando estas métricas é que a dimensionalidade é significativamente reduzida, portanto, reduzindo a complexidade do modelo de classificação. Além disso, usando somente palavras (*tokens*) como atributos pode gerar um grande número de atributos mesmo para pequenos conjuntos de dados. No método proposto é possível classificar *posts* com alta acurácia usando somente 24 atributos. Embora os atributos gerados não tenham significado linguístico, a intenção é que eles devem capturar informações latentes e propriedades semânticas do texto enquanto o SRI obtém propriedades léxicas. Os atributos extraídos dispensam, por exemplo, a presença de pontuação e emoticons.

Em nossa hipótese, *posts* similares tendem a compartilhar a mesma polaridade do sentimento. Considerando o exemplo da Tabela 5.1, esta evidência não é observada considerando somente a quantidade, onde $count_- < count_+$. A quantidade neste caso é insuficiente, pois ignora a posição dos *posts* no *ranking*. Considerando as métricas derivadas de ϕ esta evidência pode ser identificada, pois considera a posição que os *posts*

Tabela 5.1: Exemplo de *ranking* para um *post* q

Posição	Escore	Classe
1	25,897	-
2	25,541	+
3	25,158	-
4	22,729	-
5	22,204	-
6	21,576	-
7	21,468	-
8	21,426	+
9	21,011	+
10	20,985	+
11	20,666	-
12	20,666	-
13	20,546	+
14	20,546	+
15	20,435	-
16	20,070	-
17	20,032	-
18	20,025	+
19	19,974	+
20	19,962	+
21	19,934	+
22	19,902	+
23	19,889	+
24	19,779	-
25	19,771	+

Fonte: Os Autores

Tabela 5.2: Atributos derivados do *ranking*

Atributo	Valor
avg_-	21,723
max_-	25,897
min_-	19,779
sum_-	260,68
$count_-$	12
ϕ_-	8,5822
$\phi_{positional_-}$	188,44
ϕ_{avg_-}	0,3950
ϕ_{max_-}	0,3313
ϕ_{min_-}	0,4338
ϕ_{sum_-}	0,0329
ϕ_{count_-}	0,7151
avg_+	20,732
max_+	25,541
min_+	19,771
sum_+	269,51
$count_+$	13
ϕ_+	5,5743
$\phi_{positional_+}$	115,51
ϕ_{avg_+}	0,2688
ϕ_{max_+}	0,2182
ϕ_{min_+}	0,2819
ϕ_{sum_+}	0,0206
ϕ_{count_+}	0,4287
$class$	-

Fonte: Os Autores

estão no *ranking*, mesmo sendo em menor quantidade, onde $\phi_- > \phi_+$.

5.4 Classificação dos *posts*

O segundo passo nesta do método envolve o uso de um modelo de aprendizagem de máquina supervisionado. Utilizando os *posts* do conjunto de treinamento rotulado como consulta, o modelo de aprendizagem é gerado ao final da Etapa de treinamento. O objetivo é obter um modelo que generalize e classifique instâncias não rotuladas. A classificação dos *posts* ocorre sobre o conjunto de testes, onde o modelo de aprendizagem busca prever o rótulo de uma instância desconhecida.

Tabela 5.3: Descrição dos atributos derivados do *ranking*

Atributo	Descrição
avg_c	Média dos escores de <i>posts</i> da classe c
max_c	Máximo escore de <i>posts</i> da classe c
min_c	Mínimo escore de <i>posts</i> da classe c
sum_c	Soma dos escores de <i>posts</i> da classe c
$count_c$	Contagem de <i>posts</i> da classe c
ϕ_c	Utiliza a posição absoluta e a posição relativa dos <i>posts</i> da classe c para calcular a proximidade do topo do <i>ranking</i>
$\phi_{positional_c}$	Utiliza a posição absoluta, a posição relativa, e o escore dos <i>posts</i> da classe c para calcular sua proximidade do topo do <i>ranking</i>
ϕ_{avg_c}	Razão entre a proximidade do topo e a média dos escores de <i>posts</i> da classe c
ϕ_{max_c}	Razão entre a proximidade do topo e o máximo escore de <i>posts</i> da classe c
ϕ_{min_c}	Razão entre a proximidade do topo e o mínimo escore de <i>posts</i> da classe c
ϕ_{sum_c}	Razão entre a proximidade do topo e a soma dos escores de <i>posts</i> da classe c
ϕ_{count_c}	Razão entre a proximidade do topo e a contagem de <i>posts</i> da classe c

Fonte: Os Autores

Embora não exista um modelo eficaz para todas as tarefas de classificação, aplicamos os atributos gerados a diversos algoritmos de classificação como forma de identificar qual obtém maior identidade com os atributos extraídos. Classificadores de diversas famílias como funções (*Logistic Regression*, *Simple Logistic Regression*, e *Support Vector Machines*), árvores (*J48* e *REPTree*), bayesianos (*Naive Bayes* e *Multinomial Naive Bayes*), e baseados em regras (*Conjunctive Rule*) foram testados.

5.5 Metodologia de avaliação

Como forma de contribuir para o estudo da atribuição de polaridade para *posts* e complementar este trabalho, nesta seção serão apresentados os experimentos executados sobre dois conjuntos de dados reais amplamente utilizados na literatura. Estes resultados também serão comparados às *baselines* e abordagens disponíveis no estado da arte. Nas próximas seções subseções são descritas os detalhes dos experimentos e os resultados obtidos. Adicionalmente, apresentamos uma discussão sobre o efeito da escolha do número

de documentos recuperados no *ranking* (n) e como influencia nos resultados.

Conjuntos de dados. Dois conjuntos de dados de *tweets* reais, que também foram utilizados em outros trabalhos Bakliwal et al. (2002), Coletta et al. (2014), Go, Bhayani e Huang (2009), Saif, He e Alani (2012), Silva, Hruschka e Jr. (2014), Speriosu et al. (2011), foram utilizados nos experimentos. Tais como:

- Stanford-Twitter Sentiment Corpus (STD)² que é dividido em dois subconjuntos *STD-train* e *STD-test*. *STD-train* é composto de 1,6 milhões de *tweets* automaticamente rotulados. O método de rotulação automática é conhecido como *noisy-labelling* (GO; BHAYANI; HUANG, 2009), e consiste basicamente em assinar um sentimento ao *tweet* baseado no sentimento associados a emoções presentes no *tweet*. Por exemplo: *tweets* contendo emoticons como :), :-), =(, :D são classificados como positivos e *tweets* contendo :(, :'(, :-(), =(são considerados como negativos. Este método elimina o esforço manual na rotulação, mas resulta em um conjunto rotulado com ruído. *Tweets* com ironia e sarcasmo, por exemplo, podem ser classificados incorretamente. *STD-test*, por outro lado, tem 359 *tweets* que foram rotulados manualmente. Os *tweets* foram obtidos usando a API do Twitter usando consultas sobre diversos tópicos como nomes de produtos, empresas e pessoas. A Tabela 5.4 apresenta os detalhes do conjunto *STD-test*.
- Health Care Reform (HCR)³ é um conjunto de *tweets* que foi coletado em 2010. Ele contém *tweets* com a hashtag "#hcr" que significa a reforma na saúde dos Estados Unidos e foi um assunto bastante controverso na época. Os *tweets* foram manualmente anotados como positivos, negativos e neutros⁴. A Tabela 5.5 mostra os detalhes do conjunto HCR.

Pré-processamento. Os textos dos *tweets* foram modificados da seguinte forma: todos os caracteres foram convertidos para minúsculo, caracteres não-alfabéticos (incluindo números e pontuação) foram removidos, e caracteres repetidos foram substituídos por somente uma ocorrência. Todos os *tweets* utilizados (incluindo treinamento e teste) foram pre-processados.

Sistema de Recuperação de Informação. Existem diversos SRI disponíveis para utilização. Neste trabalho optou-se por utilizar o sistema Zettair Billerbeck et al. (2004) devido sua simplicidade e desempenho para indexação. Okapi BM25 foi utilizado para gerar o

²disponível em <http://help.sentiment140.com>

³Disponível em <https://bitbucket.org/speriosu/updown>

⁴*Tweets* neutros não foram considerados em nos experimentos

Tabela 5.4: Detalhes do *dataset* STD

Categoria	#Tweets	#Positivos	#Negativos
Company	119	33	86
Misc	67	26	41
Person	65	48	17
Product	63	47	16
Movies	19	16	3
Location	18	4	14
Events	8	8	0
Total	359	182	177

Fonte: Adaptado de Go, Bhayani e Huang (2009)

Tabela 5.5: Detalhes do *dataset* HCR

<i>Dataset</i>	#Tweets	#Positivos	#Negativos
HCR- <i>test</i>	658	154	504
HCR- <i>train</i>	614	213	401
Total	1.272	367	905

Fonte: Adaptado de Speriosu et al. (2011)

ranking para todas as consultas. O escore de similaridade entre *tweets* q_i e t_j é calculado como:

$$BM25(q_i, t_j) = \sum_{w \in q_i} \log \left(\frac{m - f_w + 0.5}{f_w + 0.5} \right) \times \frac{(k_1 + 1)f_{w,t}}{K + f_{w,t}} \quad (5.3)$$

onde w é uma palavra no *tweet* (*post*), m é o número de *tweets* indexados, f_w é o número de ocorrências da palavra w no *dataset*, e $f_{w,t}$ é o número de ocorrências de w em t_j . K é $k_1((1 - b) + b \times L_t/avgL)$ onde L_t é o tamanho de t_j em bytes e $avgL$ é a média de tamanho dos *tweets* indexados. Para os valores das constantes k_1 e b foram utilizados 1,2 e 0,75, respectivamente. Estes são os valores padrão para o Zettair e não foram alterados. Nenhum experimento sobre o impacto destas constantes nos resultados foi realizado.

A indexação foi aplicada somente a conjunto STD-*train* (que são *tweets* rotulados automaticamente). O índice gerado pelo Zettair tem 1.596.295 documentos (ou *tweets*), um total de 20.768.630 *tokens* (ou palavras) das quais 178.518 são distintas. A indexação destes documentos levou 8 segundos em um processador i5 3.2 GHz com 8GB de memória. O índice foi consultado usando os *tweets* dos conjuntos STD-*test*, HCR-*train*, e

HCR-*test* para calcular os atributos. Como motivação de campanhas de avaliação como TREC e CLEF, para cada consulta, os primeiros mil (1000) *tweets* mais similares foram recuperados ($n = 1000$). O tempo total utilizado para consultar o índice foi de 10 segundos. Uma análise do impacto de n nos resultados a será discutido na seção 5.6.1).

Ferramenta de classificação. Foram utilizados os classificadores implementados no Weka Hall et al. (2009) de acordo com seus parâmetros padrão. Os resultados mostram que os melhores algoritmos variam de acordo com o conjunto de dados. Embora os melhores resultados tenham sido obtidos por *Naïve Bayes Multinomial* (para STD) e *Maximum Entropy* (para HCR), foi escolhido o *Simple Logistic* como referência nos resultados por estar entre os 3 melhores resultados em ambos os *datasets*. *Simple Logistic* (LANDWEHR; HALL; FRANK, 2005) é um classificador que utiliza modelos regressão logística linear. Para comparar os resultados obtidos com os apresentados na literatura, utilizou-se validação cruzada com 10 sub-conjuntos (*10-fold cross validation*) para o *dataset* STD-*test*, enquanto que para o *dataset* HCR, o treinamento foi executado sobre o conjunto HCR-*train* e o teste sobre HCR-*test*.

Definição dos *baselines*. Foram treinados quatro classificadores tipicamente utilizados na literatura: *Naïve Bayes* (NB), *Naïve Bayes Multinomial* (MNB), *Support Vector Machine* (SVM), e *Máxima Entropia* (MaxEnt). Como atributos, foram utilizadas todas as *tokens* (unigramas) contidas no *dataset*.

5.6 Resultados

A Tabela 5.6 apresenta os valores das métricas definidas na Seção 2.4 (Acurácia, Precisão, Revocação e Medida-F) para nossos experimentos e para os *baselines*. Resultados publicados na literatura também estão incluídos. Em comparação com os *baselines* usando unigramas como atributos, foram obtidos diferenças significativas sobre o melhor *baseline* (MNB). Foi realizado um teste de Wilcoxon Signed-Rank⁵ sobre os resultados de acurácia obtendo *p-value* de 0,85 e 0,20 para STD e HCR, respectivamente. O método desenvolvido foi estatisticamente superior a MaxEnt (*p-value* de 0,0019) considerando o conjunto HCR. Em todos os outros casos, marcados com , não há diferença significativa entre o método e os *baselines*. Entretanto, nota-se que esses resultados similares foram obtidos utilizando somente 24 atributos, enquanto que os *baselines* utilizam 11.683

⁵Este teste foi usado pois os dados não seguem uma distribuição normal, o que nos impede de usar testes paramétricos como o Teste-T.

e 4.623 atributos nos *datasets* STD e HCR, respectivamente.

Em relação aos melhores resultados publicados na literatura, a o método desenvolvido obteve resultado inferior na acurácia do que a reportada por Bakliwal et al. (2002) sobre o conjunto STD em aproximadamente sete pontos percentuais. No entanto, esses resultados foram obtidos sobre 1,6 milhões de instancias de treinamento, além de ferramentas de processamento custosas como correção de texto, analisadores gramaticais, e dicionários.

A Tabela 5.6 também mostra o número de atributos (#Att) utilizado em cada um destes trabalhos⁶. Muitos dos trabalhos utilizam unigramas Coletta et al. (2014), Go, Bhayani e Huang (2009), Saif, He e Alani (2012), Saif et al. (2015), Silva, Hruschka e Jr. (2014) ou unigramas e bigramas Speriosu et al. (2011) como atributos, resultando em alguns milhares de atributos. Quando compara-se os resultados obtidos com outros trabalhos que tentam reduzir o número de atributos, nosso método superou a *feature hashing* (SILVA; HRUSCHKA; JR., 2014) no conjunto STD, mas não no *dataset* HCR, embora o método desenvolvida utilize um modelo de classificação simples enquanto que Silva, Hruschka e Jr. (2014) utiliza uma combinação de diversos modelos de classificação. Na Tabela 5.6, um \diamond indica que estatisticamente não existe diferença entre o método desenvolvido e o *baseline*, enquanto que um * indica que há diferença significativa.

Tabela 5.6: Resultado para atribuição de polaridade em *posts*

Método	STD					HCR				
	#Att	Acc	Rec	Pr	F1	#Att	Acc	Rec	Pr	F1
NB	1,6K	75,76 \diamond	75,77	75,84	75,73	4,6K	73,23 \diamond	73,23	74,71	73,88
MNB	1,6K	80,77 \diamond	80,78	80,78	80,78	4,6K	79,09 \diamond	79,10	76,50	75,92
SVM	1,6K	74,65 \diamond	74,65	74,66	74,64	4,6K	75,03 \diamond	75,79	77,05	75,85
MaxEnt	1,6K	78,55 \diamond	78,55	78,55	78,55	4,6K	67,96*	67,97	76,87	70,34
Método Desenvolvido	24	80,22	80,22	80,24	80,22	24	76,69	76,69	73,28	73,85
Bakliwal et al. (2002)	7	87,20	-	-	-	-	-	-	-	-
Coletta et al. (2014)	>1,7K	81,84	-	-	81,85	>4,6K	79,62	-	-	75,06
Go, Bhayani e Huang (2009)	364K	82,70	-	-	-	-	-	-	-	-
Silva, Hruschka e Jr. (2014)All	2K	81,06	81,10	81,09	81,00	1,4K	76,99	74,12	74,85	75,53
Silva, Hruschka e Jr. (2014)FH	21	79,11	77,17	77,19	77,15	10	78,35	77,61	74,72	75,16
Saif, He e Alani (2012)	>37K	-	83,80	84,25	83,90	>2K	-	67,40	69,15	68,15
Saif et al. (2015)	~1,7K	80,33	-	-	77,52	~4,6K	70,68	-	-	58,89
Speriosu et al. (2011)	~6K	84,7	-	-	-	~20K	71,20	-	-	-
Carvalho, Prado e Plastino (2014)	~6K	77,20	77,20	78,10	77,90	~20K	75,50	75,50	68,30	71,70

Fonte: O autor

Análise de Erros. Para identificar padrões em erros de classificação foram analisados casos onde a suposição de que *posts* similares compartilham a mesma classe de sentimento

⁶Em alguns casos, quando o trabalho não menciona o número exato, foi apresentada uma estimativa baseada na descrição disponível, e foram incluídas as notações \sim ou $>$ para indicar que o valor não é exato.

falhou. Considerando estes casos, foram estabelecidas três principais motivos: (i) *posts* interrogativos que contém palavras de sentimento, mas que não expressam sentimentos; (ii) presença de ironia; (iii) *posts* onde negações invertem a polaridade do sentimento. Em todos estes casos, os problemas podem ser atribuídos a anotação automática sem intervenção humana. Considerando que não foram utilizadas ferramentas de PLN para tratar estes casos, eles podem afetar os resultados de classificação.

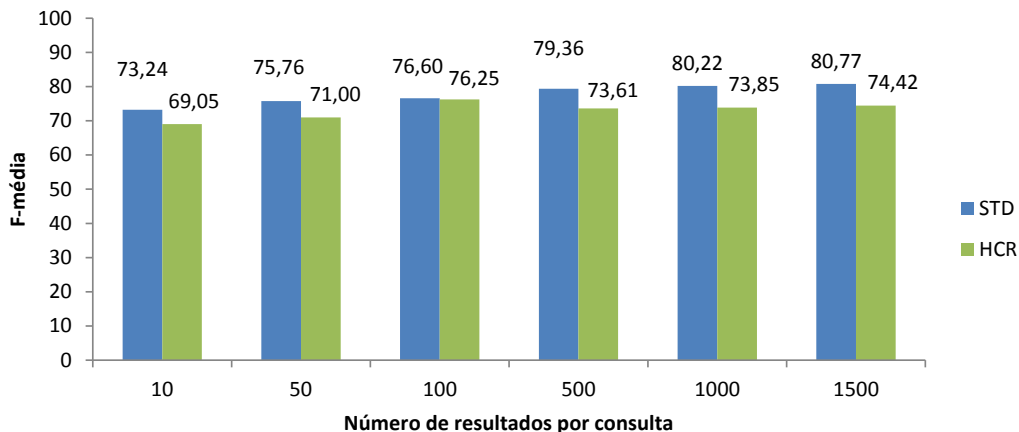
Tempo de Processamento. O tempo de indexação do *dataset* com 1,6 milhões de *posts* é de 8 segundos. O tempo necessário para processar todas as consultas foi de 5 segundos. Para treinar-se um classificador com BOW como atributos, o tempo variou entre 0.01 e 94 segundos (dependendo do algoritmo). Já o tempo de treinamento de um classificador com os 24 atributos propostos foi no máximo 17 segundos.

5.6.1 Sensibilidade de Medida-F em função de n

O método desenvolvido necessita um número de documentos recuperados n como parâmetro de entrada. Intuitivamente, n não pode ser muito pequeno pois não forneceria quantidade de evidência necessária. Por outro lado, se n for muito alto, *tweets* que não são similares à consulta seriam recuperados adicionado ruído aos atributos. Para avaliar este impacto, foram testados diversos valores de n 10, 50, 100, 500, 1000, e 1500. Dado que o índice gerado possui cerca de 1,6 milhões de *tweets*, a quantidade recuperada representa uma pequena fração do total (aproximadamente 0.006%). Estes resultados são apresentados na Figura 5.2. Os melhores resultados em termos de F1 são obtidos quando $n = 1500$ usando o *dataset* STD e $n = 100$ para HCR. Ainda assim, a maior variação, considerando todos os resultados foi de sete pontos percentuais, o que mostra que nossa estratégia é sensível a este parâmetro e a definição do valor ideal é uma oportunidade para estudos futuros.

5.6.2 Análise dos atributos

Com a finalidade de analisar o potencial discriminativo dos atributos, foi calculado o InfoGain (detalhado na Seção 2.3.2) sobre os dados de treinamento. Os valores em termos de InfoGain apresentados na Tabela 5.7 são notavelmente maiores no conjunto STD. Considerando as métricas de classificação na Tabela 5.6, nota-se que no conjunto

Figura 5.2: Sensibilidade de F1 em função de n 

Fonte: O autor

HCR foram obtidos os piores resultados, seja na literatura, *baselines* ou em nosso método. Isso sugere que este conjunto seja mais difícil de classificar e pode explicar o pior escore de InfoGain. Além disso, os melhores atributos são completamente diferentes entre STD e HCR. Entretanto nota-se que os melhores atributos em ambos os casos incluem diversos tipos atributos de contagem a métricas mais elaboradas. Isto mostra que para melhor generalização do modelo a utilização de diversos tipos de atributos é uma estratégia importante para a generalização do modelo.

Tabela 5.7: Melhores atributos segundo o InfoGain

STD		HCR	
IG	Feature	IG	Feature
0,382	ϕ_{avg+}	0,061	$avg+$
0,361	$count_-$	0,040	avg_-
0,361	$count_+$	0,034	ϕ_{min_-}
0,360	ϕ_{count_-}	0,034	sum_+
0,343	ϕ_{min+}	0,033	ϕ_{sum_-}
0,339	ϕ_{count+}	0,033	ϕ_{sum+}
0,332	ϕ_+	0,031	ϕ_{avg_-}
0,329	ϕ_-	0,031	$\phi_{positional+}$
0,310	$\phi_{positional_-}$	0,030	min_-
0,302	ϕ_{max+}	0,028	min_+

Fonte: O autor

5.7 Considerações Finais

Neste capítulo foi apresentado o método desenvolvido para a classificação de sentimentos em *posts* utilizando um SRI. O objetivo foi obter informações sobre a classe de *tweets* (rotulados) similares para classificar *posts* não-rotulados. O método extrai 24 atributos a partir do *ranking* gerado por um motor de busca como resposta a consulta composta pelo *tweet* que deseja-se classificar.

Em comparação com as *baselines*, os resultados são promissores. Não há diferença estatisticamente significativa entre a acurácia do método desenvolvido e a dos *baselines* executados nos mesmos *datasets*. Em comparação com os resultados do estado da arte, foram obtidos resultados competitivos utilizando apenas 24 atributos propostos.

6 SUMARIZAÇÃO DE OPINIÕES

Nos Capítulos 4 e 5 foram apresentadas as principais atividades desenvolvidas neste trabalho. Neste Capítulo, apresentamos a atividade final que combina ambas as tarefas já discutidas (Análise baseada em Aspectos e Atribuição de Polaridade). Apesar destas tarefas serem fundamentais para a compreensão dos sentimentos, seus resultados precisam ser apresentadas de forma simples e direta.

A sumarização de opiniões visa criar sumários que resumam as informações processadas por um sistema de análise de opiniões. Assim, como forma de complementar os trabalhos desenvolvidos, apresentamos duas formas de sumários. A primeira, resume a informação direta sobre as categorias processadas e a segunda oferece uma visualização comparativa dos resultados obtidos com o resultado esperado.

6.1 Sumarização de Opiniões sobre Categorias

Como descrito no Capítulo 2 um sumário resume as informações processadas por um sistema de análise de opiniões. Além disso, é útil para fins de comparação entre produtos similares, por exemplo. Nesta seção apresentamos, na forma de sumários, os resultados obtidos nos experimentos comparados aos resultados esperados.

Para tanto, seguiu-se a seguinte metodologia: para os *datasets* de revisões de produtos (descritos na Seção 4.4.1), foram desconsideradas opiniões neutras e o valor de avaliação é o percentual de positivos dividido pelo total de opiniões em uma categoria; para os *datasets* de *tweets*, foi considerados somente o *dataset* STD (descrito na Seção 5.5) pois ele é o único que contém informação relativa às categorias. Da mesma forma, o valor de avaliação é o percentual de positivos dividido pelo total de opiniões em uma categoria.

A Figura 6.1a apresenta a visualização do sumário de opiniões para o *dataset* Restaurantes. Nota-se grande proximidade das avaliações principalmente para as categorias *FOOD*, *RESTAURANT* e *SERVICE* que juntas compõem a maioria absoluta das opiniões classificadas.

A Figura 6.1b apresenta a Visualização do Sumário de opiniões para o *dataset* Laptops. Embora este data-set tenha um maior número de categorias esparsas, nota-se que as avaliações estão aproximadas, principalmente para as categorias com maior número de opiniões como *LAPTOP*, *MEMORY* e *DISPLAY* sendo que o inverso também é frequente. Nestes casos, atribuímos os erros de classificação a quantidade limitada de

instâncias de treinamento. Visto que quantidades menores tendem a ser mais específicas e considerando o fato de utilizarmos um método genérico para atribuição de polaridade.

A Figura 6.1c apresenta a Visualização do Sumário de opiniões para o *dataset* Oculto. Este *dataset* foi o maior desafio para o método, sendo que o domínio é desconhecido. Assim o sumário apresenta avaliações díspares para a maioria das categorias, a exceção foi a categoria *FOOD_DRINKS* que se aproximou do resultado esperado.

A Figura 6.1d apresenta a Visualização do Sumário de opiniões para o *dataset* STD. A atribuição de polaridade em *tweets* é bastante complexa devido a diversidade de informações. Embora para este *dataset* estejam agrupados em algumas categorias facilite a análise sobre determinados pontos de erros. Os resultados obtidos reportam mais de 80% de F-média, entretanto somente este valor é insuficiente para apontar pontos de ajustes. Por exemplo, as categorias *Company*, *Person* e *Product* representam os maiores pontos de erros, enquanto que as categorias *Misc.* e *Movie* estão mais consistentes com os resultados esperados.

Para todos os *datasets* analisados, é possível obter ainda mais detalhes através de sumários que apresentem também as informações de subcategorias. Entretanto, resulta em um grande volume de informação que dificulta o entendimento do sumário.

6.2 Sumarização de Opiniões Comparativa

A visualização de opiniões é importante principalmente por facilitar o entendimento de grandes quantidades de dados. Neste trabalho utilizamos os conceitos de sumarização para expor os resultados obtidos nos experimentos comparados aos resultados esperados (Gold). Para o entendimento do sumário utilizamos algumas convenções: primeira, a intensidade do sentimento é observada no eixo Y, onde -1 representa negativos e 1 representa positivos; segunda, somente opiniões positivas e negativas foram consideradas; e terceira, a quantidade de opiniões não escala com a quantidade real para fins de visualização. A Figura 6.2a apresenta os resultados obtidos para a classificação de opiniões para o *dataset* Restaurantes, discutido no Capítulo 4.

Em nossos experimentos sobre classificação de revisões no *dataset* Restaurantes, foram obtidos 71,72% de acurácia. Esta informação pode ser claramente observada na Figura 6.2a, onde se observa a similaridade dos resultados obtidos comparados aos resultados esperados, principalmente para a categoria *SERVICE*. Observa-se também uma grande discrepância para a categoria *LOCATION*

A Figura 6.2b apresenta a visualização dos resultados de classificação no *dataset* Laptops. Nota-se grande proximidade entre os resultados obtidos e esperados, principalmente para as categorias *FANS_COOLING*, *GRAPHICS*, *OPTICAL_DRIVES*, *POWER_SUPPLY* e *SHIPPING*. No entanto, observa-se a dificuldade em classificar sentenças para as categorias *BATTERY* e *MEMORY*. Para este domínio, nosso método obteve 67,33% de acurácia.

A Figura 6.2c apresenta os resultados para o *dataset* Oculito. Para este *dataset* não havia dados de treinamento para a identificação de polaridade, assim nossa estratégia foi útil para este cenário. Entretanto, nota-se as diferenças de classificação em todas as categorias. Isto resultou no pior resultado de classificação dentre os três domínios (65,78%).

A visualização de resultados de classificação de *tweets* tende a ser mais difícil, pois na maioria dos casos a única informação que se tem sobre eles é sua polaridade. Neste trabalho, foi utilizado o *dataset* STD que foi criado usando determinadas categorias de consultas (GO; BHAYANI; HUANG, 2009). Utilizamos estas informações para apresentar os erros de classificação de nosso método. A Figura 6.2d apresenta os resultados obtidos para cada uma das categorias de consulta. Esta visualização também apresenta informações importantes sobre como foram classificados os tweets em suas categorias. Os resultados obtidos de acurácia (80,22%) são bastante consistentes com os resultados esperados.

Figura 6.1: Visualização de sumários de opiniões baseados em aspectos

(a) Visualização do Sumário de opiniões para o *dataset* Restaurantes

Categoria	Resultado Esperado			Resultado Experimento		
	Positivos	Negativos	Avaliação	Positivos	Negativos	Avaliação
AMBIENCE	51	20	★★★★★	43	29	★★★★★
DRINKS	13	9	★★★★★	10	12	★★★★★
FOOD	220	103	★★★★★	188	136	★★★★★
LOCATION	4	0	★★★★★	6	2	★★★★★
RESTAURANT	119	91	★★★★★	97	95	★★★★★
SERVICE	47	123	★★★★★	42	118	★★★★★

(b) Visualização do Sumário de opiniões para o *dataset* Laptops

Categoria	Resultado Esperado			Resultado Experimento		
	Positivos	Negativos	Avaliação	Positivos	Negativos	Avaliação
BATTERY	15	6	★★★★★	12	10	★★★★★
COMPANY	7	10	★★★★★	3	13	★★★★★
CPU	5	6	★★★★★	6	5	★★★★★
DISPLAY	38	8	★★★★★	36	9	★★★★★
FANS_COOLING	0	2	★★★★★	0	1	★★★★★
GRAPHICS	0	2	★★★★★	0	2	★★★★★
HARD_DISC	2	2	★★★★★	1	3	★★★★★
HARDWARE	2	2	★★★★★	1	3	★★★★★
KEYBOARD	20	6	★★★★★	16	9	★★★★★
LAPTOP	414	222	★★★★★	302	293	★★★★★
MEMORY	4	0	★★★★★	3	2	★★★★★
MOUSE	5	29	★★★★★	8	26	★★★★★
MULTIMEDIA_DEVICES	9	6	★★★★★	8	8	★★★★★
OPTICAL_DRIVES	0	4	★★★★★	0	4	★★★★★
OS	11	13	★★★★★	10	16	★★★★★
POWER_SUPPLY	0	1	★★★★★	0	1	★★★★★
SHIPPING	1	0	★★★★★	1	0	★★★★★
SOFTWARE	6	7	★★★★★	2	10	★★★★★
SUPPORT	2	4	★★★★★	1	5	★★★★★

(c) Visualização do Sumário de opiniões para o *dataset* Oculto

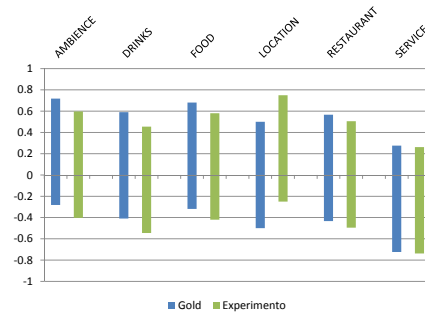
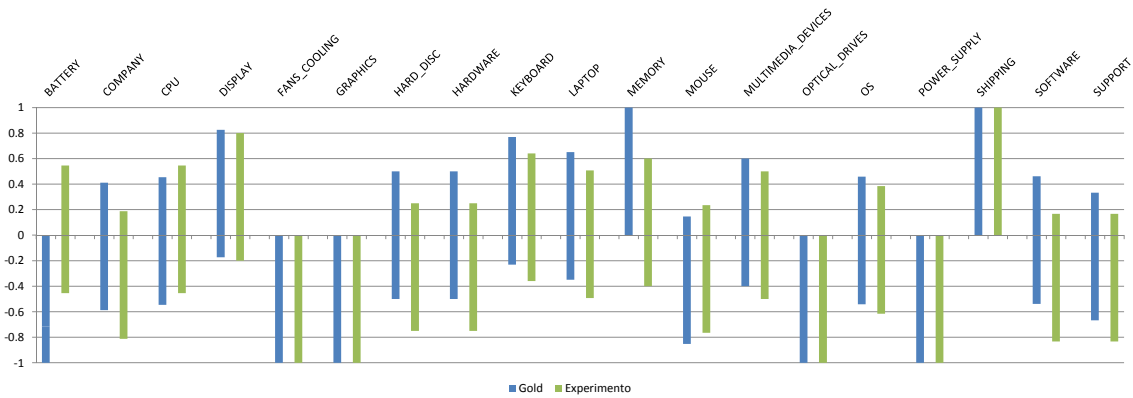
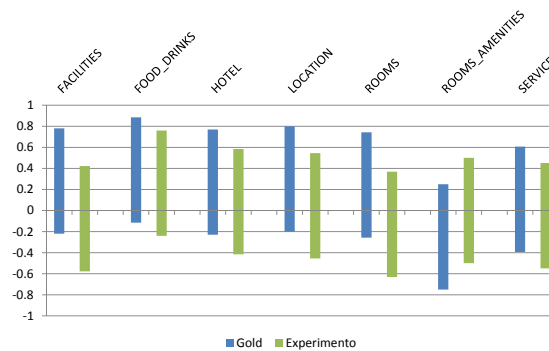
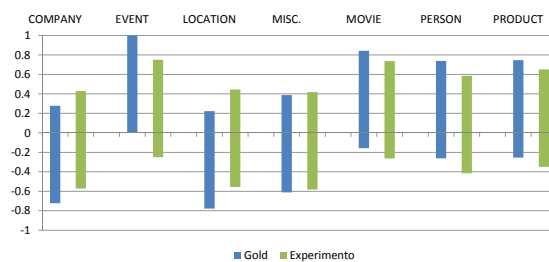
Categoria	Resultado Esperado			Resultado Experimento		
	Positivos	Negativos	Avaliação	Positivos	Negativos	Avaliação
FACILITIES	39	11	★★★★★	19	26	★★★★★
FOOD_DRINKS	23	3	★★★★★	19	6	★★★★★
HOTEL	77	23	★★★★★	52	37	★★★★★
LOCATION	20	5	★★★★★	12	10	★★★★★
ROOMS	49	17	★★★★★	24	41	★★★★★
ROOMS_AMENITIES	1	3	★★★★★	2	2	★★★★★
SERVICE	34	22	★★★★★	23	28	★★★★★

(d) Visualização do Sumário de opiniões para o *dataset* STD

Categoria	Resultado Esperado			Resultado Experimento		
	Positivos	Negativos	Avaliação	Positivos	Negativos	Avaliação
Company	33	86	★★★★★	51	68	★★★★★
Event	8	0	★★★★★	6	2	★★★★★
Location	4	14	★★★★★	8	10	★★★★★
Misc.	26	41	★★★★★	28	39	★★★★★
Movie	16	3	★★★★★	14	5	★★★★★
Person	48	17	★★★★★	38	27	★★★★★
Product	47	16	★★★★★	41	22	★★★★★

Fonte: O autor

Figura 6.2: Visualização de sumários comparativos de opiniões

(a) Visualização do Sumário comparativo para o *dataset* Restaurantes(b) Visualização do Sumário comparativo para o *dataset* Laptops(c) Visualização do Sumário comparativo para o *dataset* Oculto(d) Visualização do Sumário comparativo para o *dataset* STD

Fonte: O autor

7 CONCLUSÃO

Este Capítulo conclui a dissertação apresentando uma síntese das ideias propostas no decorrer do trabalho, as contribuições e os resultados obtidos. Também serão apresentadas as limitações e deficiências da metodologia proposta e, finalmente, as direções apontadas para trabalhos futuros.

7.1 Resumo das contribuições

Neste trabalho foram desenvolvidas duas tarefas principais no contexto de Análise de Sentimento: a identificação de aspectos e alvos de opinião e a atribuição de polaridade. A contribuição sobre identificação de aspectos e alvos da opinião dá-se pela redução do uso de recursos externos ao conjunto de treinamento e redução da dimensionalidade.

Atualmente, métodos para a identificação de aspectos estão cada vez mais dependentes de recursos externos ao treinamento, como dicionários de sentimentos e/ou combinação de inúmeras ferramentas de processamento. Os resultados obtidos superaram as *baselines* utilizando uma ferramenta de processamento de linguagem natural e nenhum recurso externo ao conjunto de treinamento. Embora para a identificação de alvos tenhamos utilizado uma lista de exceções, em nossos experimentos, provou-se que o ganho em seu uso não influencia nos resultados finais. Quanto à redução da dimensionalidade, apresentamos o método de classificação em duas fases que cria modelos genéricos para categorias e reduz significativamente o número de atributos de classificação. Além disso, os resultados obtidos superam os da classificação em uma fase, que utiliza somente informações textuais.

A contribuição sobre a identificação de polaridade está nos atributos utilizados e na redução da dimensionalidade. O método desenvolvido utiliza um motor de busca para obter os *posts* mais similares àquele que se quer classificar. Em nossa hipótese, consideramos que *posts* similares tendem a compartilhar a mesma polaridade de sentimento. A partir do ranking gerado, são extraídos 24 atributos que capturam características latentes do texto. Os atributos são utilizados por algoritmos de classificação para gerar modelos de aprendizagem. Os experimentos realizados atingiram resultados melhores do que *baselines* que utilizam o modelo *bag-of-words* e grande parte do estado da arte. Além disso, o uso de um número pequeno de atributos representa uma economia de recursos computacionais. O método desenvolvido ainda apresenta duas vantagens: não depende

de recursos linguísticos e pode ser aplicado sobre dados com ruídos.

7.2 Limitações e trabalhos futuros

Algumas das limitações desse trabalho, que motivam a continuação do mesmo, são apresentadas a seguir.

1. Para a análise de sentimentos baseada em aspectos, utilizamos um método para a seleção de palavras candidatas. Neste método, a escolha do limiar ($InfoGain \geq 0$) pode ter impactado negativamente nos resultados. Uma direção para próximos estudos é explorar alternativas para a seleção destas palavras sementes, tais como a utilização de métodos de aprendizagem não-supervisionados. Entretanto, isso não evitaria a necessidade de definir outros parâmetros, assim a melhor alternativa é estudar a sensibilidade do método para este parâmetro.
2. Neste trabalho, a identificação de categorias utiliza classificação binária e independente para cada categoria presente no domínio. Esta tarefa pode ser explorada como um problema de classificação multi-rótulo visando, por exemplo, identificar padrões de coocorrências entre categorias.
3. Para a atribuição de polaridade, também utilizamos um limiar que corresponde ao número de documentos recuperados por consulta. Embora tenhamos constatado que a variação deste valor não impacta significativamente os resultados, a limitação neste ponto está diretamente relacionada ao tamanho do índice do motor de busca. Um possível segmento de estudo está na análise do impacto do tamanho do índice nos resultados.
4. O índice utilizado pelo motor de busca contém dados com ruído. Embora os resultados sejam comparáveis com apontados na literatura, a acurácia máxima fica próxima a 80%. Uma alternativa interessante, seria o desenvolvimento de métodos capazes de remover parte do ruído contido no índice. Neste trabalho utilizamos a métrica de similaridade Okapi BM25, outras métricas também podem ser exploradas.
5. Em ambas as contribuições, os métodos desenvolvidos são construídos sobre algoritmos de aprendizagem de máquina supervisionados. A limitação nestes casos deve-se à dependência de um conjunto genérico para treinamento.
6. Finalmente, os métodos para identificação de aspectos e atribuição de polaridade

não trabalham em conjunto. Embora as duas contribuições apresentadas sejam tratadas de forma independente, como trabalho futuro seria interessante utilizar o método de atribuição de polaridade para os aspectos extraídos.

7.3 Publicações

Durante o desenvolvimento deste trabalho foram redigidos e submetidos diversos artigos científicos relacionados ao tema da dissertação. Dentre esses, seguindo a ordem cronológica, destacam-se:

- Em 2013, desenvolvemos um método não-supervisionado para identificar tópicos em coleções de texto que foi publicado como um *short paper* no SBBD (KAUER; MOREIRA, 2013).
- Em 2014, participei do Workshop de Teses e Dissertações em Banco de Dados (KAUER, 2014) onde apresentei as ideias iniciais deste trabalho.
- Em 2015 desenvolvemos um método para a identificação de aspectos que foi submetido à campanha de avaliação SemEval 2015 (KAUER; MOREIRA, 2015), onde superou o *baseline* em todas as tarefas. Este estudo foi discutido em profundidade no Capítulo 4.
- Em 2016, a identificação de polaridade baseada em atributos extraídos de um motor de busca foi submetido ao periódico *Expert Systems with Applications* (Qualis A2).

7.4 Outros estudos desenvolvidos

Além de trabalhos direcionados à análise de sentimentos, também contribuí com trabalhos em andamento de outros membros do grupo de pesquisa. Em 2014 desenvolvemos um método para identificação de perfil de autoria baseada em textos do autor (WEREN et al., 2014). Este estudo inspirou a pesquisa de atributos gerados a partir de sistemas de recuperação de informação, principalmente pelos bons resultados sobre dados com ruído. Entre 2013 e 2015 também pesquisamos métodos para automatizar a extração de informações sobre conferências. Esse trabalho ainda está em desenvolvimento.

REFERÊNCIAS

- BAKLIWAL, A. et al. Mining sentiments from tweets. In: **WORKSHOP IN COMPUTATIONAL APPROACHES TO SUBJECTIVITY AND SENTIMENT ANALYSIS**, 3. 2012, WASSA '12, Jeju, Korea: Proceedings... Stroudsburg, USA, 2002. p. 11–18.
- BARBOSA, L.; FENG, J. Robust sentiment detection on twitter from biased and noisy data. In: **POSTER IN INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS**, 23. COLING '10, Beijing, China: Proceedings... Stroudsburg, USA, 2010. p. 36–44.
- BILLERBECK, B. et al. Rmit university at trec 2004. In: **Proceedings Text Retrieval Conference (TREC)**. [S.l.: s.n.], 2004.
- BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 993–1022, mar. 2003. Accessed 6 apr. 2016. Available from Internet: <<http://dl.acm.org/citation.cfm?id=944919.944937>>.
- CARVALHO, J.; PRADO, A.; PLASTINO, A. A statistical and evolutionary approach to sentiment analysis. In: **CONFERENCES ON WEB INTELLIGENCE (WI) AND INTELLIGENT AGENT TECHNOLOGIES (IAT)**, 2. 2014, WI-IAT '14, Washington, DC, USA: Proceedings... IEEE Computer Society, 2014. p. 110–117.
- CHANG, A. X.; MANNING, C. Sntime: A library for recognizing and normalizing time expressions. In: **Conference on Language Resources and Evaluation**. 2012, LREC '12, Istanbul, Turkey: Proceedings... European Language Resources Association (ELRA), 2012.
- CHEN, D.; MANNING, C. A fast and accurate dependency parser using neural networks. In: **CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING (EMNLP)**. 2014, EMNLP '14, Doha, Qatar: Proceedings... Association for Computational Linguistics, 2014. p. 740–750.
- COLETTA, L. F. S. et al. Combining classification and clustering for tweet sentiment analysis. In: **CONFERENCE ON INTELLIGENT SYSTEMS**. 2014, BRACIS '14, Washington, DC, USA: Proceedings... IEEE Computer Society, 2014. p. 210–215.
- CUNNINGHAM, H. et al. Gate: an architecture for development of robust hlt applications. In: **In Recent Advanced in Language Processing**. [S.l.: s.n.], 2002. p. 168–175.
- DAVIDOV, D.; TSUR, O.; RAPPOPORT, A. Enhanced sentiment learning using twitter hashtags and smileys. In: **POSTER IN CONFERENCE ON COMPUTATIONAL LINGUISTICS**. 2010, COLING '10, Stroudsburg, PA, USA: Proceedings... Association for Computational Linguistics, 2010. p. 241–249.
- DEMPSTER, A. P.; LAIRD, N. M.; RUBIN, D. B. Maximum likelihood from incomplete data via the em algorithm. **Journal of the Royal Statistical Society, Series B**, v. 39, n. 1, p. 1–38, 1977.

FERRUCCI, D.; LALLY, A. Uima: An architectural approach to unstructured information processing in the corporate research environment. **Nat. Lang. Eng.**, Cambridge University Press, New York, NY, USA, v. 10, n. 3-4, p. 327–348, sep. 2004. Accessed 6 apr. 2016. Available from Internet: <<http://dx.doi.org/10.1017/S1351324904003523>>.

FINKEL, J. R.; GRENAGER, T.; MANNING, C. Incorporating non-local information into information extraction systems by gibbs sampling. In: **ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS**, 43. 2005, ACL '05, Ann Arbor, Michigan: Proceedings... Stroudsburg, PA, USA, 2005. p. 363–370.

GHIASSI, M.; SKINNER, J.; ZIMBRA, D. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. **Expert Systems with Applications**, v. 40, n. 16, p. 6266 – 6282, 2013.

GO, A.; BHAYANI, R.; HUANG, L. Twitter sentiment classification using distant supervision. **CS224N Project Report, Stanford**, v. 1, p. 12, 2009.

HALL, M. et al. The weka data mining software: An update. **SIGKDD Explor. Newsl.**, ACM, New York, NY, USA, v. 11, n. 1, p. 10–18, nov. 2009. ISSN 1931-0145.

HOFMANN, T. Probabilistic latent semantic indexing. In: **CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL**, 22. 1999, SIGIR '99, New York, NY, USA: Proceedings... ACM, 1999. p. 50–57.

HU, M.; LIU, B. Mining and summarizing customer reviews. In: **CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING**, 10. 2004, KDD '04, New York, NY, USA: Proceedings... ACM, 2004. p. 168–177. ISBN 1-58113-888-1.

HU, X. et al. Exploiting social relations for sentiment analysis in microblogging. In: **CONFERENCE ON WEB SEARCH AND DATA MINING**, 6. 2013, WSDM '13, New York, NY, USA: Proceedings... ACM, 2013. p. 537–546.

JAKOB, N.; GUREVYCH, I. Extracting opinion targets in a single- and cross-domain setting with conditional random fields. In: **CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING**. 2010, EMNLP '10, Stroudsburg, PA, USA: Proceedings... Association for Computational Linguistics, 2010. p. 1035–1045.

JIN, W.; HO, H. H. A novel lexicalized hmm-based learning framework for web opinion mining. In: **CONFERENCE ON MACHINE LEARNING**. 2009, ICML '09, New York, NY, USA: Proceedings... ACM, 2009. p. 465–472.

KAUER, A. Mineração de opinião multilíngue para comentários sobre produtos. In: **WORKSHOP ON THESIS AND DISSERTATIONS IN DATABASES BRAZILIAN SYMPOSIUM ON DATABASES**, 29. 2014, SBBB '14, Curitiba, Paraná: Proceedings... SBC, 2014. p. 325–331.

KAUER, A.; MOREIRA, V. Ufrgs: Identifying categories and targets in customer reviews. In: **WORKSHOP ON SEMANTIC EVALUATION**, 9. 2015, SemEval '15, Denver, Colorado: Proceedings... Association for Computational Linguistics, 2015. p. 725–729.

KAUER, A.; MOREIRA, V. P. Evolução dos temas de interesse do sbdd ao longo dos anos. In: **POSTER ON BRAZILIAN SYMPOSIUM ON DATABASES**). 2014, SBBD '14, Recife, Brazil: Proceedings... SBC, 2013.

KLEIN, D.; MANNING, C. D. Fast exact inference with a factored model for natural language parsing. In: **Advances in Neural Information Processing Systems**. MIT Press, 2003. v. 15. Accessed 6 apr. 2016. Available from Internet: <<http://www-nlp.stanford.edu/~manning/papers/lex-parser.pdf>>.

LAFFERTY, J. D.; MCCALLUM, A.; PEREIRA, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: **CONFERENCE ON MACHINE LEARNING**. 2001, ICML '01, San Francisco, CA, USA: Proceedings... Morgan Kaufmann Publishers Inc., 2001. p. 282–289.

LANDWEHR, N.; HALL, M.; FRANK, E. **Logistic Model Trees**, Kluwer Academic Publishers, Hingham, MA, USA, v. 59, n. 1-2, p. 161–205, may 2005. Accessed 6 apr. 2016. Available from Internet: <<http://dx.doi.org/10.1007/s10994-005-0466-3>>.

LEE, H. et al. Deterministic coreference resolution based on entity-centric, precision-ranked rules. **Comput. Linguist.**, MIT Press, Cambridge, MA, USA, v. 39, n. 4, p. 885–916, dec. 2013. Accessed 6 apr. 2016. Available from Internet: <http://dx.doi.org/10.1162/COLI_a_00152>.

LITA, L. V. et al. tRuEcasIng. In: **ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS**,41. 2003, ACL '03, Stroudsburg, PA, USA: Proceedings... Association for Computational Linguistics, 2003. p. 152–159.

LIU, B. Sentiment analysis and opinion mining. **Synthesis Lectures on Human Language Technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

LIU, B.; HU, M.; CHENG, J. Opinion observer: analyzing and comparing opinions on the web. In: **WWW**. [S.l.: s.n.], 2005. p. 342–351.

LIU, Q. et al. A logic programming approach to aspect extraction in opinion mining. In: **Web Intelligence (WI) and Intelligent Agent Technologies (IAT)**. [S.l.: s.n.], 2013. p. 276–283.

LOPER, E.; BIRD, S. Nltk: The natural language toolkit. In: **WORKSHOP ON EFFECTIVE TOOLS AND METHODOLOGIES FOR TEACHING NATURAL LANGUAGE PROCESSING AND COMPUTATIONAL LINGUISTICS**,2. 2002, ETMTNLP '02, Philadelphia, Pennsylvania: Proceedings... Stroudsburg, PA, USA, 2002. p. 63–70.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZ, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

MANNING, C. D. et al. The Stanford CoreNLP natural language processing toolkit. In: **ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: SYSTEM DEMONSTRATIONS**,52. 2014, San Francisco, CA, USA: Proceedings... Association for Computational Linguistics, 2014. p. 55–60.

MARNEE, M.-C. de; MANNING, C. D. Stanford typed dependencies manual. 2008. Accessed 6 apr. 2016. Available from Internet: <<http://ufal.mff.cuni.cz/~hladka/2015/docs/Stanford-dependencies-manual.pdf>>.

MARTÍNEZ-CÁMARA, E. et al. Sentiment analysis in twitter. **Natural Language Engineering**, v. 20, p. 1–28, 1 2014.

MITCHELL, T. M. **Machine learning**. [S.l.]: McGraw-Hill, 1997. (McGraw Hill series in computer science). ISBN 978-0-07-042807-2.

MOGHADDAM, S.; ESTER, M. Opinion digger: an unsupervised opinion miner from unstructured product reviews. In: **CIKM**. [S.l.: s.n.], 2010. p. 1825–1828.

MUKHERJEE, A.; LIU, B. Modeling review comments. In: **ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS**, 50. 2012, ACL '12, Stroudsburg, PA, USA: Proceedings... Association for Computational Linguistics, 2012. p. 320–329.

PONTIKI, M. et al. Semeval-2015 task 12: Aspect based sentiment analysis. In: **WORKSHOP ON SEMANTIC EVALUATION**, 9. 2015, SemEval '15, Denver, Colorado: Proceedings... Association for Computational Linguistics, 2015. p. 486–495.

QIU, G. et al. Opinion word expansion and target extraction through double propagation. **Computational Linguistics**, MIT Press, v. 37, n. 1, p. 9–27, 2011.

RABINER, L. A tutorial on hidden markov models and selected applications in speech recognition. **Proceedings of the IEEE**, v. 77, n. 2, p. 257–286, Feb 1989.

RODRIGUES, J. et al. Multi-entity polarity analysis and detection of subjectivity in financial documents. **Journal of Information and Data Management**, v. 6, n. 2, p. 130, 2016.

SAIAS, J.; RAMALHO, R. R. Sentiue: Target and aspect based sentiment analysis in semeval-2015 task 12. In: **WORKSHOP ON SEMANTIC EVALUATION**, 9. 2015, SemEval '15, Denver, Colorado: Proceedings... Association for Computational Linguistics, 2015. p. 767–771.

SAIF, H.; HE, Y.; ALANI, H. Semantic sentiment analysis of twitter. In: **The Semantic Web–ISWC 2012**. [S.l.: s.n.], 2012. p. 508–524.

SAIF, H. et al. Contextual semantics for sentiment analysis of twitter. **Information Processing & Management**, 2015.

SANTORINI, B. **Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)**. [S.l.], 1990. Accessed 6 apr. 2016. Available from Internet: <http://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports>.

SILVA, N. F. da; HRUSCHKA, E. R.; JR., E. R. H. Tweet sentiment analysis with classifier ensembles. **Decision Support Systems**, v. 66, p. 170–179, 2014.

SPERIOSU, M. et al. Twitter polarity classification with label propagation over lexical links and the follower graph. In: **WORKSHOP ON UNSUPERVISED LEARNING IN NLP**, 1. 2011, EMNLP '11, Edinburgh, Scotland: Proceedings... Stroudsburg, PA, USA, 2011. p. 53–63.

TOH, Z.; SU, J. Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In: **WORKSHOP ON SEMANTIC EVALUATION,9**. 2015, SemEval '15, Denver, Colorado: Proceedings... Association for Computational Linguistics, 2015. p. 719–724.

TOUTANOVA, K. et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: **CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS ON HUMAN LANGUAGE TECHNOLOGY**. 2003, NAACL '03, Edmonton, Canada: Proceedings... Stroudsburg, PA, USA, 2003. p. 173–180.

WEREN, E. R. et al. Examining multiple features for author profiling. **Journal of Information and Data Management**, v. 5, n. 3, p. 266, 2014.

YU, J. et al. Aspect ranking: Identifying important product aspects from online consumer reviews. In: **ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES,49**. 2011, HLT '11, Portland, Oregon: Proceedings... Stroudsburg, PA, USA, 2011. p. 1496–1505.

ZHANG, L. et al. **Combining Lexicon-based and Learning-based Methods for Twitter Sentiment Analysis**. [S.l.], 2011. Accessed 6 apr. 2016. Available from Internet: <<http://www.hpl.hp.com/techreports/2011/HPL-2011-89.html>>.