# Minicurso
# 2

# Introdução à Mineração de Opiniões: Conceitos, Aplicações e Desafios

Karin Becker, Diego Tumitan

Programa de Pós-Graduação em Ciência da Computação - Instituto de Informática
Universidade Federal do Rio Grande do Sul (UFRGS)
{karin.becker,dctumitan}@inf.ufrgs.br

***Abstract***

*Social networks, foruns, microblogs, on-line newspapers and sites for the evaluation of products and services are examples of the many platforms available in the web, which allow users to express their opinions. Opinion mining, or sentiment analysis, is a recent field that aims at identifying opinionative content, and determine the sentiment, perception or attitude of the public with regard to the target of the opinion. This Chapter discusses the motivation for the field and its main applications; presents the issues of opinion mining and related concepts; describes a set of techniques for extracting opinions and their target, classifying their sentiment, and summarizing opinions; and concludes with challenges and research perspectives for this field.*

***Resumo***

*Redes sociais, fóruns, tweets, jornais on-line, sites para avaliação de produtos e serviços, são alguns exemplos de plataformas através das quais usuários têm expresso suas ideias e opiniões. A mineração de opiniões, ou análise de sentimentos, é uma disciplina recente que busca identificar conteúdo de opinião, e determinar o sentimento, percepção ou atitude do público em relação ao alvo desta opinião. Este capítulo discute a motivação para a área e suas principais aplicações; apresenta o problema de mineração de opiniões e conceitos relacionados; descreve um conjunto de técnicas para extrair opiniões e seu alvo, classificar seu sentimento ou polarizá-las, e sumarizar as opiniões; e conclui com desafios e perspectivas de pesquisa na área.*

## 1.1. Introdução

Opiniões têm grande influência sobre o comportamento das pessoas. Decisões simples como qual carro comprar, qual filme ver, ou em qual ação investir são frequentemente baseadas em opiniões de pessoas próximas, de especialistas, ou de estudos conduzidos por instituições especializadas. Organizações baseiam suas estratégias de negócio e investimentos na opinião de seus clientes sobre seus produtos ou serviços. A importância da opinião é tão grande que muitas empresas (e.g. marketing, relações públicas, pesquisas) têm seu negócio voltado à obtenção deste tipo de informação. Tradicionalmente, a resposta a questões envolvendo a opinião pública envolve técnicas como pesquisa de campo, telefonemas ou questionários escritos. Estas técnicas envolvem custos, são restritas a um grupo focal bem definido ou amostra, seu retorno é demorado e muitas vezes, pouco eficaz. A latência da opinião também é alta, devido ao longo tempo necessário entre a coleta dos dados brutos, sua análise e disponibilização dos resultados.

A explosão das mídias sociais alterou este cenário, disponibilizando a indivíduos e organizações conteúdo de opinião diversificado e em grandes volumes. Usuários da web têm a oportunidade de registrar e divulgar suas ideias e opiniões através de comentários, fóruns de discussão, blogs, Twitter, redes sociais, entre outros. Isto aumenta as opções dos indivíduos na busca de opiniões, pois não estão mais limitados a sua rede pessoal de contatos (e.g. familiares, amigos, conexões profissionais) ou a opiniões de especialistas disponíveis publicamente (e.g. revistas, jornais). No tocante às organizações, isto significa oportunidades de ampliar as fontes de opinião quantitativa ou qualitativamente, tornar mais baratas as formas de coleta, e reduzir o tempo necessário para disponibilização da informação. Contudo, o grande volume de informação produzido diariamente implica a necessidade de métodos e ferramentas capazes de processar automaticamente não apenas o conteúdo das publicações, mas também a opinião e sentimento que expressam.

A *mineração de opiniões*, também chamada de *análise de sentimento* ou *análise de subjetividade* [20, 34, 23], é uma disciplina recente que congrega pesquisas de mineração de dados, linguística computacional, recuperação de informações, inteligência artificial, entre outras. A mineração de opinião é definida em [19] como qualquer estudo feito computacionalmente envolvendo opiniões, sentimentos, avaliações, atitudes, afeições, visões, emoções e subjetividade, expressos de forma textual. O problema da mineração de opiniões pode ser estruturado em termos das seguintes tarefas genéricas [34]: a) identificar as opiniões expressas sobre determinado assunto ou alvo em um conjunto de documentos; b) classificar a orientação ou polaridade desta opinião, isto é, se tende a positiva ou negativa; e c) apresentar os resultados de forma agregada e sumarizada. A polaridade da opinião define o sentimento, percepção ou atitude do público em relação ao alvo da opinião.

Como mostra a Figura 1.1, boa parte dos trabalhos nesta área concentra-se no desenvolvimento de técnicas para detecção e sumarização automáticas de opinião sobre revisões de produtos e serviços [17, 24, 37, 12, 15, 2]. Posteriormente, o foco ampliou-se para entidades específicas (e.g. políticos, celebridades, marcas) em redes sociais [22, 10] ou notícias [16]. Várias ferramentas foram desenvolvidas com este objetivo, entre elas

TweetSentiments[1], Sentimonitor[2], Google Shopping[3], UberVU[4], etc. A mineração de opinião sobre textos menos estruturados, como notícias e blogs, também tem sido alvo de bastante atenção [6, 7, 18]. Outra vertente importante é a análise de opinião do Twitter, em particular visando estabelecer modelos preditivos [3, 8, 35].

O propósito deste capítulo é apresentar os conceitos subjacentes à mineração de opiniões, e caracterizar cada uma das etapas do processo, descrevendo os problemas envolvidos, e as técnicas que podem ser utilizadas. Etapas, problemas e técnicas serão ilustrados através de trabalhos representativos encontrados na literatura, considerando diferentes tipos de texto: revisões de produtos, notícias e mídias sociais.

O restante deste capítulo está estruturado como segue. Na Seção 1.2 é descrita a fundamentação teórica, com o detalhamento dos conceitos que caracterizam a opinião, dos diferentes níveis de análise de opinião, e dos problemas de análise textual relacionados. A Seção 1.3 descreve o processo de mineração de opiniões em termos de suas principais etapas: identificação, classificação de polaridade e sumarização. As diferentes abordagens para a classificação de polaridade são apresentadas na Seção 1.4. A mineração de opiniões sobre diferentes tipos de dados é então discutida através de exemplos na Seção 1.5, considerando revisões de produtos, notícias e mídias sociais. Finalmente, a Seção 1.6 apresenta conclusões e direções futuras.

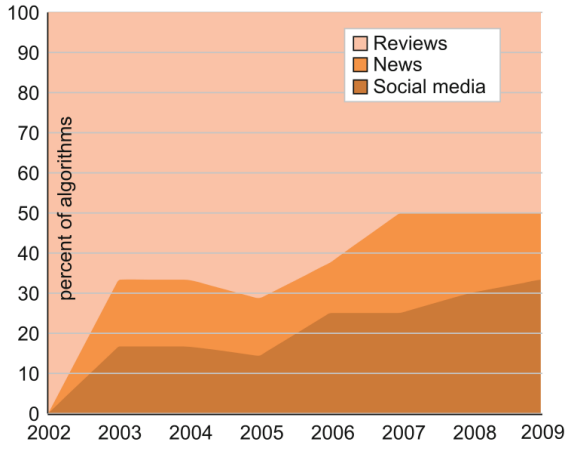


**Figura 1.1. Distribuição dos trabalhos sobre domínios alvo (Fonte: [34]).**

## 1.2. Conceitos

### 1.2.1. Definições

A mineração de opiniões opera sobre porções de texto de quaisquer tamanho e formato, tais como páginas web, posts, comentários, *tweets*, revisões de produto, etc. Toda opinião é composta de pelo menos dois elementos chave: um *alvo* e um *sentimento* sobre este alvo [20]. Um alvo pode ser uma entidade, aspecto de uma entidade, ou tópico, representando um produto, pessoa, organização, marca, evento, etc. Já um sentimento representa

[1]http://twittersentiment.appspot.com
[2]http://www.sentimonitor.com
[3]http://www.google.com/shopping
[4]http://www.ubervu.com

uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. A *polaridade* de um sentimento corresponde a um ponto em alguma escala que representa a avaliação positiva, neutra ou negativa do significado deste sentimento [34].

Mais formalmente, uma opinião corresponde a uma quíntupla $(\mathbf{e}_i, \mathbf{a}_{ij}, \mathbf{s}_{ijkl}, \mathbf{h}_k, \mathbf{t}_l)$ [19], onde:

- $\mathbf{e}_i$: é o nome de uma entidade;
- $\mathbf{a}_{ij}$: é um aspecto da entidade $\mathbf{e}_i$ (opcional);
- $\mathbf{s}_{ijkl}$: é a polaridade do sentimento sobre aspecto $\mathbf{a}_{ij}$ que tem como alvo a entidade $\mathbf{e}_i$;
- $\mathbf{h}_k$: é o detentor do sentimento (i.e. quem expressou o sentimento), também chamado de fonte de opinião;
- $\mathbf{t}_l$: é o instante no qual a opinião foi expressa por $\mathbf{h}_k$.

O conceito de *aspecto*, também denominado característica (*feature*) ou propriedade, permite que uma entidade seja vista através de diferentes perspectivas ou atributos, ou como uma hierarquia de partes e subpartes [19]. Por exemplo, considere o comentário abaixo sobre um hotel, retirado do *site* Booking.com. Os aspectos deste hotel são o quarto, a vista, e o *wi-fi*. Assim, o sentimento expresso neste comentário é representado por quatro quíntuplas, sendo que uma refere-se ao hotel, e as demais, a aspectos específicos deste.

- *Cláudio - 3 de setembro de 2013: "Adorei o hotel **Vida Mansa**. Os **quartos** do hotel são super espaçosos, com uma **vista** linda para o mar. Pena que não há **wi-fi** nos quartos".*
  – (Vida Mansa, geral, positivo, Cláudio, 03/09/2013)
  – (Vida Mansa, quarto, positivo, Cláudio, 03/09/2013)
  – (Vida Mansa, vista, positivo, Cláudio, 03/09/2013)
  – (Vida Mansa, wi-fi, negativo, Cláudio, 03/09/2013)

Os termos *sentimento* e *opinião* frequentemente são usados como sinônimo neste contexto. A polaridade de um sentimento pode ser classificada em classes discretas (e.g. positiva, negativa ou neutra), ou como um intervalo que representa a intensidade deste sentimento, tipicamente [-1, 1]. Já o termo *emoção* é usado para designar as percepções e pensamentos subjetivos de uma pessoa, tais como raiva, desgosto, medo, alegria, tristeza e surpresa, não representando necessariamente um posicionamento ou uma atitude em relação ao alvo.

### 1.2.2. Níveis de Análise Textual

A detecção do sentimento em um texto pode ocorrer em diferentes granularidades, sendo que a decisão do nível está sujeita ao contexto e aplicação. A análise pode ser em nível de [20]:

- *Documento*: nesse nível, a tarefa é classificar se um documento, tratado como um todo, expressa um sentimento positivo ou negativo. Esta granularidade é adequada quando o documento trata de uma única entidade, por exemplo, um documento que forneça uma opinião sobre um dado produto;

- *Sentença*: determina o sentimento de uma sentença específica de um documento. Este nível é bastante utilizado quando um mesmo documento contém opiniões sobre várias entidades. Ele também permite identificar e distinguir sentenças objetivas (fatos) e subjetivas (opiniões). Alguns autores sugerem ir além do nível de sentença, dividindo-a em cláusulas (e.g. "A cidade é péssima, mas a população é muito simpática") [33];

- *Entidade e Aspecto*: este nível foca na opinião expressa, independentemente dos construtos utilizados para expressá-la (e.g. documentos, sentenças, orações). Neste caso, o alvo da opinião pode ser uma entidade, ou algum de seus aspectos. No exemplo "Adoro minha câmera X porque a qualidade de sua lente é excepcional. Pena que o preço seja tão alto", observa-se que existem três opiniões em 2 sentenças: sobre a câmera X, e sobre dois de seus aspectos (preço e lente). Apenas a opinião sobre o preço é negativa, sendo que a opinião sobre a lente, e sobre a câmera em geral são positivas. Este é o nível mais complexo de análise, o qual tem sido bastante estudado no contexto de revisões de produtos e serviços (e.g. [17, 33, 15]).

### 1.2.3. Tipos de Opiniões e Análise Linguística

Opiniões referem-se a conteúdo subjetivo, escrito em linguagem natural. A forma como as opiniões estão expressas influencia diretamente a habilidade de processá-las corretamente. A mineração de opiniões tem origens em comum com a linguística computacional, com a qual compartilha problemas e desafios [20].

Opiniões podem ser *regulares* ou *comparativas*; *diretas* ou *indiretas*, e *implícitas* ou *explícitas*. Em opiniões regulares, o autor da opinião expressa um sentimento, atitude, emoção ou percepção sobre um alvo ("Este filme é muito bom"). Já as opiniões comparativas expressam o sentimento com base na relação de similaridades ou diferenças entre duas ou mais entidades, ou preferência quanto a algum aspecto compartilhado ("O teclado deste telefone é muito melhor do que o do meu telefone antigo"). As opiniões podem ser diretas ("Este suco é muito bom"), ou indiretas ("minha gripe piorou depois que tomei este remédio" – implicando opinião negativa sobre o remédio através do seu efeito sobre a gripe). Finalmente, opiniões explícitas expressam diretamente o sentimento, enquanto que as implícitas sugerem-no indiretamente ("Formou-se um vale no colchão que comprei na semana passada"). A maioria dos trabalhos concentra-se em opiniões regulares, diretas e explícitas, por serem mais fáceis de serem tratadas.

Um problema enfrentado é a co-referência, onde diferentes tipos de menções designam a uma mesma entidade. Por exemplo, as expressões "Dilma", "Presidenta", "Presidenta Dilma Rousseff" referem-se à mesma pessoa, devendo ser reconhecidas e unificadas. Nesse mesmo contexto, outra dificuldade é a resolução de pronomes, com o objetivo de relacionar um pronome a uma determinada entidade. Por exemplo, no texto "Paris é uma cidade maravilhosa. Ela é um excelente lugar para se visitar. Seus restaurantes são

muito reconhecidos", os pronomes "ela" e "seus" referem-se a Paris. O tratamento da co-referência e dos pronomes é extremamente importante para a análise de sentimentos nos níveis de sentença e de aspectos, já que estes níveis analisam o sentimento de forma isolada (i.e. cada sentença ou opinião), com efeito direto sobre a revocação.

É comum o uso de palavras de sentimento (e.g. ótimo, péssimo) para detectar opiniões, mas seu uso não é condição necessária, nem suficiente, para detectar uma opinião, e classificar sua polaridade. Primeiro, as palavras de opinião podem ser positivas ou negativas de acordo com o contexto. Por exemplo, na sentença "este smartphone é muito caro", a palavra de sentimento "caro" é negativa, enquanto que em "este amigo me é muito caro", ela é positiva. Segundo, nem toda opinião é expressa com palavras de sentimento (e.g. "comprei este casaco na semana passada, e já está cheio de bolinhas"), ou vice-versa (e.g. "se encontrar um bom livro, vou lê-lo"). A negação é outra questão que deve ser tratada, já que inverte o sentido da opinião ("Este filme não é nada bom").

Finalmente, a ironia/sarcasmo é um dos problemas mais difíceis de se tratar ("Ontem vendo o horário político, vi propostas bem novas e inovadoras: melhorar a saúde, educação e emprego. Por que ninguém havia prometido isto antes?"). O uso de sarcasmo é muito comum em alguns domínios, como discussões políticas e esportivas, opiniões sobre arte (filmes, bandas), etc [27, 37, 5]. Os trabalhos encontrados na literatura para identificação de sarcasmo/ironia fazem uso de artifícios, como: frequência do sinal de exclamação e interrogação, palavras capitalizadas, interjeições (e.g. "ah, oh, yeah"), *emoticons* (e.g. ";-)") e superlativos [11, 20].

## 1.3. Etapas da Mineração da Opinião

A mineração de opinião pode ser caracterizada em termos de três grandes tarefas [34]: a) identificar (tópicos, sentenças opinativas), b) classificar a polaridade do sentimento, e c) sumarizar. Este processo é esboçado na Figura 1.2.

### 1.3.1. Identificação

Dado um conjunto de textos extraídos de alguma fonte (e.g. jornais, redes sociais, plataformas de revisão de produtos/serviços), a etapa de *identificação* consiste em encontrar os tópicos existentes (e possivelmente seus aspectos), e possivelmente associá-los com o respectivo conteúdo subjetivo. A forma de identificar as entidades, aspectos e sentimento são dependentes da granularidade escolhida para análise, e os algoritmos utilizados podem ser distintos daqueles propostos para recuperação de documentos opinionativos [23, 34].

A complexidade da identificação do alvo da opinião depende em muito da mídia considerada, e de seu grau de estruturação. A aplicação mais frequente em mineração de opiniões é a de revisão de produtos e serviços, porque o alvo pode ser mais facilmente identificado. Assume-se que todo o documento refere-se a uma única entidade, o alvo da revisão, sendo que o desafio está em identificar os aspectos desta entidade, se a análise for nesta granularidade.

Já em jornais, blogs ou posts, não se conhece *a priori* as entidades envolvidas, podendo inclusive envolver muitas entidades na mesma porção de texto. Na situação mais simples, pode-se restringir a identificação a entidades pré-definidas, como a busca

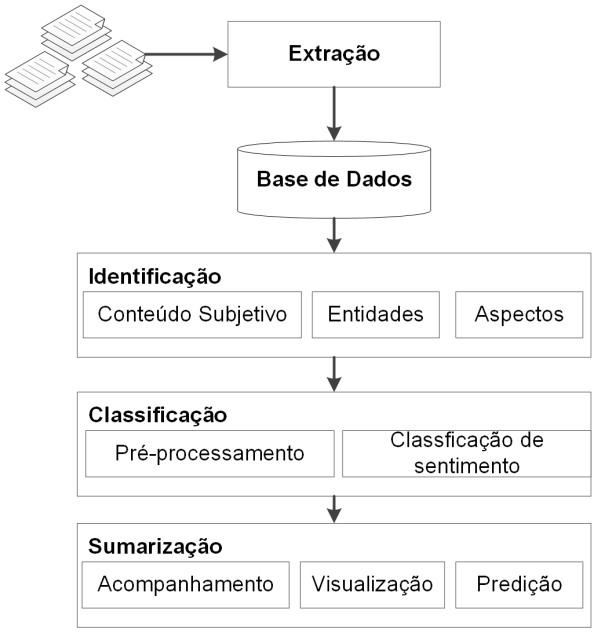


**Figura 1.2. Etapas da Mineração de Opinião.**

de celebridades, atletas, políticos ou marcas. Um dos problemas neste caso é resolver os problemas de co-referência e de pronomes, já mencionados na Seção 1.2.3. Em mídias sociais, a co-referência pode ser um problema acentuado, pois as menções podem ser muito informais (apelidos, gírias com significado local ou temporal, *hashtags*, etc). Por exemplo, o termo "tricolor" no estado de São Paulo refere-se ao São Paulo Futebol Clube, enquanto que no estado do Rio Grande do Sul, esse termo designa o Grêmio Foot-Ball Porto Alegrense. Se a identificação não for direcionada a alvos pré-definidos, pode-se ainda utilizar técnicas de identificação de entidades nomeadas da recuperação de informações [26, 1].

Finalmente, esta tarefa pode envolver também o discernimento entre conteúdo ou sentenças com ou sem opinião, visando melhorar os resultados da próxima etapa. Isto é bastante comum quando o nível de análise é de granularidade menor. O critério utilizado para determinar o conteúdo de opinião é quase sempre a identificação de palavras de sentimento (e.g. "Eu recomendo este filme"), ou de classes de palavras candidatas a expressar sentimento (e.g. adjetivos).

### 1.3.2. Classificação da Polaridade

O problema de *classificação de sentimento*, também denominado *classificação de polaridade*, é frequentemente um problema de classificação binário, isto é, que classifica um dado texto em uma de duas classes: *positivo* ou *negativo*. No entanto, classes adicionais podem ser consideradas para que a análise seja mais robusta, ou para aumentar o nível de detalhe dos resultados. Assim, estas classes podem ser desdobradas em classificações com diferentes graus de intensidade (e.g. muitoPositivo, moderadamentePositivo), ou em intervalos numéricos representando um grau de intensidade [34]. Neste último caso, a divisão do sentimento está relacionada à capacidade de definir algum limiar para distinguir os níveis de sentimento.

Outra abordagem é considerar a categoria neutra, que engloba textos sem uma tendência clara quanto a sua polaridade, ou simplesmente sem sentimento. Neste último caso, é a etapa de classificação de polaridade que tem como responsabilidade identificar textos sem sentimento de acordo com suas propriedades. Mas, como já mencionado, é mais frequente que este tipo de texto já tenha sido descartado na etapa anterior de identificação, porque a qualidade dos resultados da classificação costuma ser maior [34].

Para a classificação da polaridade, diferentes abordagens são propostas na literatura, as quais são discutidas com maiores detalhes na Seção 1.4. Cada técnica pode necessitar de operações de pré-processamento e transformação específicas, tais como reconhecimento de construtos sintáticos, reconhecimento de n-gramas, extração de *features*, eliminação de termos irrelevantes, transformação em vetor de termos, etc.

Independente da abordagem empregada, a classificação da polaridade não é um problema trivial. Entre os principais desafios estão:

- o uso de palavras de sentimento pode ser enganoso, como discutido na Seção 1.2.3, já que existem opiniões sem o uso de palavras de sentimento, e vice-versa. Ainda, a polaridade de alguns termos são dependentes de contexto;
- muitos domínios são caracterizados pelo uso frequente de ironias ou sarcasmo, onde o sentido implícito é exatamente oposto ao sentimento expresso explicitamente. Outros domínios (e.g. debates políticos, críticas culturais) estabelecem uma opinião positiva por oposição a uma argumentação negativa (ou vice-versa) [5, 24, 37];
- a opinião pode depender do observador. Por exemplo, a opinião representada na sentença "As ações da Petrobrás subiram" é positiva para quem detém este tipo de ação, mas pode ser péssima para quem deixou de investir nelas;
- a polaridade de conteúdo subjetivo nem sempre é objeto de consenso. Por exemplo, em anotações feitas por humanos, dificilmente o consenso é maior que 75% [9, 18, 39, 24];
- a classificação é bastante dependente da extração das *features* do texto, a qual deve lidar com as várias questões da língua natural já discutidas na Seção 1.2.3.

É importante ressaltar que um texto neutro é diferente de um texto não polarizado. Um texto não polarizado é aquele no qual não há elementos suficientes para poder classificá-lo, e consequentemente, para o qual a tarefa de classificação não consegue chegar à conclusão sobre sua polaridade. Isso geralmente acontece quando o conteúdo analisado possui ruídos, tais como erros tipográficos e sentenças incompletas [13].

### 1.3.3. Sumarização

Para poder identificar opinião média ou prevalecente de um grupo de pessoas sobre um determinado tópico/entidade, a opinião expressa por uma única pessoa não é suficiente, sendo necessário analisar uma grande quantidade de opiniões [20]. É necessário a criação de métricas e sumários que quantifiquem a diversidade de opiniões encontradas a respeito

um mesmo alvo. Este é o objetivo desta etapa, onde são criadas métricas que representem o sentimento geral, as quais podem ser visualizadas ou servir de entrada para outras aplicações.

Em revisão de produtos, um sumário de um determinado produto pode ajudar um consumidor a identificar seus respectivos pontos fortes e fracos, levando em consideração a experiência prévia de outras pessoas, expressas em suas opiniões. Esse tipo de recurso pode ser encontrado, por exemplo, no Google Shopping, que automaticamente extrai, analisa e agrega os aspectos de revisões de produtos disponibilizados por diferentes lojas de comércio eletrônico (e.g. Best Buy). A Figura 1.3 ilustra um exemplo de resultado desta ferramenta, onde aspectos como facilidade de uso, *design* e tamanho, foram identificados e exemplificados com uma sentença que demonstra o sentimento predominante sobre o mesmo. Além disso, a ferramenta ainda mostra uma nota geral sobre o produto, baseada em uma classificação de estrelas.

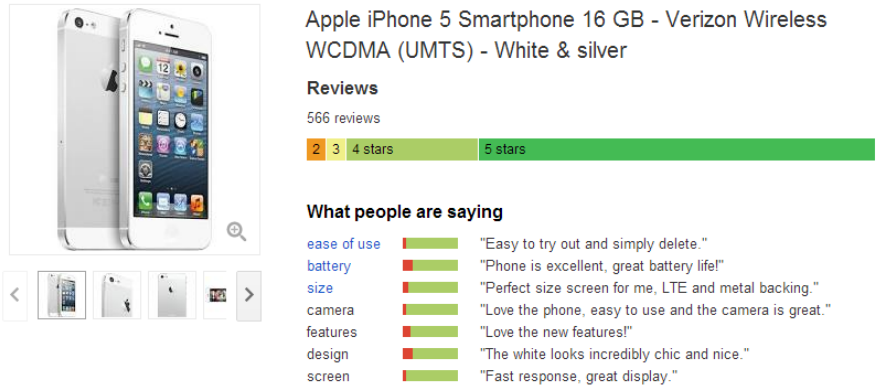


**Figura 1.3. Exemplo de sumarização de opiniões de aspectos produto extraídas de revisões (Fonte: Google Shopping).**

Outra forma de sumarização, comum em aplicações que extraem de mídias sociais o sentimento do público em geral sobre uma determinada entidade (e.g. uma marca, produto, político, celebridade), é apresentar o sentimento na forma de relógios, ou associá-lo a informações temporais ou geográficas. Normalmente este tipo de mídia reflete o que as pessoas pensam sobre o alvo, dado algum evento. Por exemplo, o lançamento de um novo produto terá impacto nas redes sociais, que expressarão reações a esse acontecimento através de posts, comentários, *tweets*, endossos, etc. A empresa pode aproveitar-se disto para avaliar se este produto foi bem recepcionado pelo mercado. Um exemplo é a ferramenta UberVB, que sumariza e acompanha as menções e o sentimento em relação a uma determinada marca através do tempo. Os dados analisados são provenientes de várias mídias sociais, como o Twitter, Facebook, YouTube, blogs, etc. Na Figura 1.4 é mostrado o sentimento em relação à marca Microsoft, medindo-se o sentimento positivo, negativo e neutro através de um relógio, e sua evolução temporal.

O sentimento sumarizado também pode ser utilizado para diversas aplicações, como prever eleições [35], comportamento da bolsa de valores [8], arrecadação de bilheterias de filmes [3], definição de preços [2], etc. No entanto, o sentimento puro (positivo ou negativo) não pode refletir de maneira correta o contexto analisado. Portanto, é im-

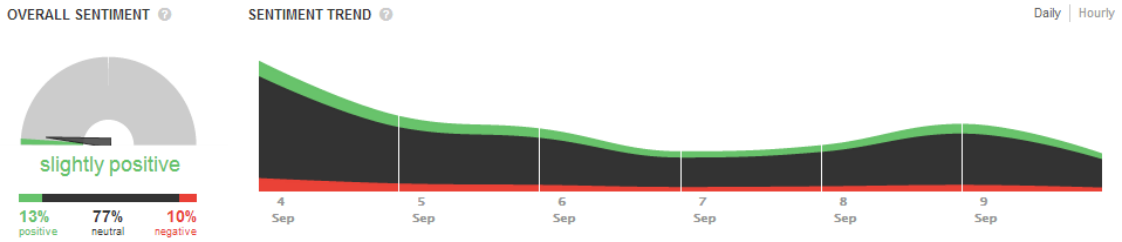


**Figura 1.4. Sentimento extraído de mídias sociais em relação à Microsoft (Fonte: UberVU).**

portante criar métricas para representar o sentimento em relação ao alvo. Boa parte dos trabalhos na área utilizam a média do sentimento, ou a razão entre o sentimento positivo e negativo. Em certos casos, a predição pode ser feita somente com base na quantidade de menções às entidades, independente do sentimento sobre elas (e.g. [35]).

Por exemplo, Social Market Analytics[5] é uma ferramenta para analisar a bolsa de valores, e ações de uma determinada companhia utilizando o Twitter. Na Figura 1.5 é mostrada a análise temporal do sentimento em relação às ações da *APPL - Apple Inc*. Esta ferramenta criou suas próprias métricas para análise, baseando na representação normalizada e ponderada da série de tempo do sentimento ao longo de um período retrospectivo (*S-Score*), e também uma média suavizada da métrica anterior (*S-Mean*). Estas métricas são comparadas com o valor de fechamento de uma determinada ação.

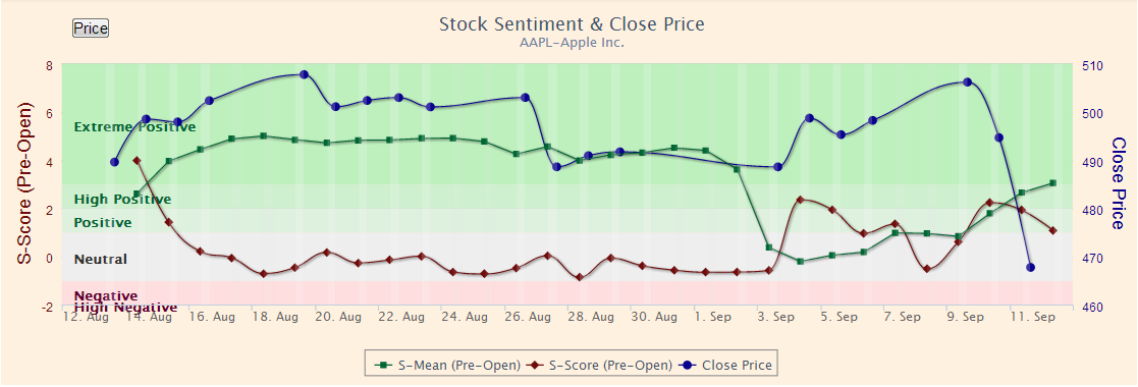


**Figura 1.5. Relação entre sentimento e preço de ações (Fonte: Social Market Analytics).**

## 1.4. Abordagens de Classificação de Polaridade

As abordagens de classificação podem ser divididas em quatro grandes grupos: a) léxicas, com o uso de dicionários de sentimentos; b) aprendizado de máquina, com o uso predominante de técnicas de classificação ou de regressão; c) estatísticas, que valem-se de técnicas para avaliar a co-ocorrência de termos, e d) semânticas, que definem a polaridade de palavras em função de sua proximidade semântica com outras de polaridade conhecidas. Técnicas destas diferentes abordagens podem ser combinadas para melhoria de resultados. Uma revisão sistemática descrita em [34] aponta uma predominâncias

[5]http://socialmarketanalytics.com

das duas primeiras abordagens, sem que nenhuma técnica se sobressaia em termos de desempenho.

### 1.4.1. Abordagem Baseada em Dicionário

A abordagem baseada em *dicionário* é também denominada *léxica* ou *linguística*. O aspecto central desta abordagem é o uso de léxicos (dicionários) de sentimentos, que são compilações de palavras ou expressões de sentimento associadas à respectiva polaridade.

Um dos métodos mais utilizados na abordagem linguística é o da co-ocorrência entre alvo e sentimento, que não leva em consideração nem a ordem dos termos dentro de um documento (*bag-of-words*), nem suas relações léxico-sintáticas. Para a classificação do sentimento em um texto, basta que exista uma palavra de sentimento, onde sua polaridade é dada por um léxico de sentimentos. Esse método é extensamente empregado para o atrelamento de um sentimento a uma entidade em uma sentença. Por exemplo, na sentença "o iPhone é muito bom", a polaridade positiva da palavra "bom" é associada à entidade iPhone. O método por co-ocorrência apresenta bons resultados quando o nível de análise textual é de granularidade pequena, pois a palavra detentora do sentimento está próxima à entidade que qualifica. Sendo assim, este método é usualmente utilizado em análises de nível de sentença, cláusula ou até em documentos com poucos caracteres, como um *tweet*.

Quando aplicada em nível de maior granularidade, estabelece-se algum tipo de média sobre as palavras de sentimento encontradas. A Equação 1 mostra uma função genérica de determinação de polaridade em um documento D, onde $S_w$ representa a polaridade de uma palavra $w$ em um dicionário. A agregação pode levar em conta funções de peso e de modificação. A função $peso()$ pode ser, por exemplo, alguma medida de distância entre a palavra de sentimento e o alvo, ou de importância da palavra no texto (e.g. frequência). A função $modificador()$ pode ser usada para tratar negações, palavras de intensidade (e.g. muito), etc. Esta função de agregação também pode ser estendida a sentenças, cujas cláusulas podem combinar diferentes palavras de sentimento.

$$\mathrm{S(D)} = \frac{\sum_{w \in D} S_w \cdot peso(w) \cdot modificador(w)}{\sum peso(w)} \tag{1}$$

Existem métodos linguísticos mais complexos, como a utilização de *parsers* linguísticos, que têm como propósito analisar o texto e aumentar a qualidade da classificação com base em informações morfossintáticas ali presentes (e.g. sujeito, predicado, dependências, funções sintáticas, etc.). No entanto, ferramentas de processamento de linguagem natural são em sua maioria restritas a determinado idioma. Recursos para a língua portuguesa são escassos, quando comparada à língua inglesa, situação esta comum a outra línguas.

A composição básica de um léxico de sentimento é a palavra de sentimento com suas possíveis flexões (e.g. bonito, bonita, bonitos), e sua respectiva polaridade, expressa como uma categoria, ou como um valor em uma escala. Muitos dicionários possuem adicionalmente: o lema e o *stem* de cada entrada; a categoria gramatical (*Part-Of-Speech - POS*); e o alvo do sentimento (predicado ou sujeito). A maioria dos léxicos existentes são dependentes de idioma e foram feitos estritamente para a língua inglesa, como General

Inquirer [30], OpinionFinder [38], SentiWordNet [4] e WordNetAffect [31]. Já para a língua portuguesa estão disponíveis o OpLexicon [29], e o SentiLex-PT [28], sendo o primeiro para português do Brasil e o último, para português de Portugal. Outro exemplo, é o Linguistic Inquiry and Word Counts (LIWC) [25], que é um software de análise de texto desenvolvido para avaliar os componentes estruturais, cognitivos e emocionais de amostras de texto, sendo que essa análise pode ser feita em vários idiomas disponibilizados na ferramenta. A Tabela 1.1 mostra uma comparação entre os léxicos mencionados.

**Tabela 1.1. Tabela comparativa de léxicos de sentimentos.**

| Dicionário | Pos | Neg | PoS | Stem | Lema | Idioma |
|---|---|---|---|---|---|---|
| General Inquirer | 1.915 | 2.291 | S | N | N | Inglês |
| OpinionFinder | 2.718 | 4.912 | S | S | N | Inglês |
| OpLexicon | 8.675 | 14.469 | S | N | N | Português |
| SentiLex-PT | 82.347 entradas | | S | N | S | Português |
| SentiWordNet | 117.659 entradas | | S | N | S | Inglês |

A maioria dos dicionários disponíveis são genéricos, ou seja, auxiliam na tarefa de classificação, independentemente do domínio dos textos sendo considerados. Entretanto, os melhores resultados obtidos na tarefa de classificação foram baseados em dicionários dependentes de contextos [17], criados a partir de palavras semente e expandidos utilizando o WordNet [14] ou tesauros. Esta abordagem também é classificada como *semântica*, e será melhor discutida na Seção 1.4.3. Finalmente, léxicos são de pouca valia quando considerados em textos gerados em mídias informais (e.g. redes sociais, *tweets*), onde expressões regionais, gírias, abreviaturas típicas da internet, etc, são fartamente empregados. A velocidade na criação de um novo vocabulário ultrapassa a capacidade de verificação de sanidade dos termos [10].

### 1.4.2. Abordagem baseada em Aprendizado de Máquina

O objetivo principal das técnicas de aprendizado de máquina é descobrir automaticamente regras gerais em grandes conjuntos de dados, que permitam extrair informações implicitamente representadas. De modo geral, as técnicas de aprendizado de máquina podem ser divididas em dois tipos: aprendizado supervisionado e aprendizado não supervisionado [32].

Na área de mineração de opiniões, nota-se um predomínio do uso de métodos supervisionados de aprendizagem, mais especificamente, *classificação* e *regressão*. Neste contexto, o problema de classificação é dividido em dois passos: (1) aprender um modelo de classificação sobre um corpus de treinamento previamente rotulado com as classes consideradas (e.g. positivo, negativo); e (2) prever a polaridade de novas porções de texto com base no modelo resultante. Dentre os algoritmos de classificação mais usados nesta área estão Support Vector Machine, Naïve Bayes, Maximum Entropy e algoritmos baseados em redes neurais [12, 8, 24, 23, 34].

A qualidade do modelo preditivo resultante da etapa de aprendizagem é medida em termos de métricas como *acurácia* (capacidade do modelo de prever corretamente), *precisão* (número de instâncias previstas corretamente em uma dada classe), ou *revocação* (número de instâncias de uma dada classe previstas na classe correta). Alguns trabalhos

obtêm precisões muito maiores na classificação da polaridade negativa, do que na positiva [27, 5, 36]. Além das dificuldades próprias ao domínio, esta situação pode ser explicada pela dificuldade em tratar ironia e sarcasmo.

Os dados de treino para a classificação/regressão correspondem a um conjunto de instâncias caracterizadas por atributos. O rótulo é denominado atributo *alvo*, enquanto que os demais são designados como atributos discriminantes ou *features*. O atributo alvo na classificação é discreto, enquanto que na regressão é numérico. Em termos de pré-processamento, é necessário extrair de cada porção de texto analisada, as *features* relevantes para a tarefa de classificação e representá-las na forma de um vetor de termos, como ilustra a Tabela 1.2.

**Tabela 1.2. Exemplo de uma entrada de classificador com vetor binário de termos.**

| gosto | produto | ruim | grande | prático | não | facilidade | uso | polaridade |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | pos |
| 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | neg |
| 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | neg |

Os tipos de *feature* mais frequentemente considerados são [24, 20, 34]:

- Palavras de sentimento: somente as palavras de sentimento de um corpus são utilizadas como *feature*. Não existe ordem entre os termos, e estes são caracterizados de forma binária, isto é, presente ou ausente no texto;
- Termos e sua frequência: são usados n-gramas (de sentimento ou não), junto com sua frequência absoluta ou relativa (e.g. TF-IDF), como peso dos termos;
- *Part-of-Speech* (POS): as classes morfológicas das palavras também podem ser usadas, em complementação às palavras de sentimento ou termos;
- Dependência sintática: as dependências sintáticas entre as palavras podem ser utilizadas, com o intuito de auxiliar na definição do alvo e fonte do sentimento.

Uma das grandes limitações no uso de aprendizado supervisionado para definição de polaridade é a necessidade de dados rotulados para treino. O desempenho destes métodos é afetado não somente pela quantidade, mas igualmente pela qualidade dos dados de treino disponíveis. Ainda, cada conjunto de treino é fortemente vinculado ao seu domínio. Nos trabalhos envolvendo revisões de produto, a classificação dada pelos usuários na forma de notas ou estrelas é utilizada como o rótulo para o texto correspondente [24, 37]. Tal facilidade não está disponível para outros domínios, implicando a necessidade de anotações manuais, as quais são trabalhosas e com alto teor de subjetividade. Nestes casos, as alternativas costumam ser os métodos léxicos, ou probabilísticos/semânticos, discutidos na próxima seção.

Como uma possível solução ao problema de anotação, foi proposto um método de classificação incremental baseado em regras [38], que utiliza regras genéricas sobre estruturas sintáticas para gerar texto subjetivo rotulado, o qual é usado para treinar classificadores e gerar novas regras mais específicas. A geração automática de texto rotulado no

domínio de política também é proposta em [27]. Esta alternativa é contudo impraticável para fontes que geram dados em *streaming*, onde além de grandes volumes de dados que devem ser analisados com baixa latência, existe uma volatilidade enorme nos tópicos e termos utilizados. Uma abordagem semi-supervisionada voltada à análise de sentimentos em tempo real [10] determina a polaridade de sentimento de *tweets*, tomando como ponto de partida opositores e apoiadores conhecidos, e inferindo através de regras de associação a propagação deste sentimento nas conexões sociais.

### 1.4.3. Abordagens Estatísticas e Semânticas

As abordagens estatísticas, por vezes denominadas *não supervisionadas* [19, 20], baseiam-se na premissa de que palavras que traduzem opiniões frequentemente são encontradas juntas no corpus dos textos. Se a palavra ocorre mais frequentemente junto a palavras positivas (negativas) no mesmo contexto, então é provável que seja positiva (negativa); já se ocorre em igual frequência, a palavra deve ser neutra. A polaridade de uma palavra desconhecida pode ser determinada calculando a co-ocorrência com uma palavra notadamente positiva (negativa), tal como "excelente" ou "péssimo". A técnica mais representativa nesta categoria é a Pointwise Mutual Information (PMI) [37].

O PMI entre dois termos quaisquer x e y é calculado segundo a Equação 2, onde Pr(x e y) é a probabilidade de co-ocorrência dos termos x e y, enquanto que Pr(x).Pr(y) é a probabilidade de co-ocorrência se são estatisticamente independentes. Esta razão é portanto a medida de grau de independência estatística entre os dois termos, e o logaritmo desta razão é a quantidade de informação ganha se os termos são observados juntos. A polaridade do sentimento de um termo x, dada pela Equação 3, é a diferença entre os valores de PMI calculados a partir de duas listas opostas: termos positivos (e.g. excelente), e termos negativos (e.g. péssimo). Na proposta original deste método [37], as co-ocorrências foram obtidas a partir da internet, utilizando uma mecanismo de busca existente na época (AltaVista).

$$PMI(x,y) = \log_2\left(\frac{Pr(x \wedge y)}{Pr(x)Pr(y)}\right) \tag{2}$$

$$\textit{PMI-IR}\ (x) = \sum_{p \in pWords} PMI(x,p) - \sum_{n \in nWords} PMI(x,n) \tag{3}$$

Outras técnicas na mesma abordagem são o Semantic-orientation Latent Semantic Analysis (SO-LSA) [37], que usa a LSA para calcular a força da associação semântica entre dois termos, através da análise estatística entre eles; e a Latent Dirichlet Allocation (LDA), muito utilizada para a extração de tópicos em textos [1]. A abordagem estatística é menos suscetível à dependência de contexto quando comparada às abordagens por dicionário e por aprendizado de máquina, mas pode ser utilizada para complementá-las [34].

A abordagem *Semântica* é bastante parecida com a estatística, exceto que a polaridade é calculada em termos de alguma medida de distância entre termos. O princípio das técnicas nesta categoria (e.g. [17, 16]) é que palavras semanticamente próximas devem ter a mesma polaridade. Por exemplo, o WordNet [14] provê diferentes relacionamen-

tos entre palavras que podem ser usadas para calcular a polaridade do sentimento, tais como sinônimos e antônimos. De forma similar à abordagem estatística, palavras semente com sentimento positivo e negativo são utilizadas como ponto de partida de um processo de comparação ou expansão, que busca determinar a polaridade dos termos. Isto pode ser feito através da mensuração da distância entre palavras usando as relações como caminho (e.g. [16]), ou da contagem da frequência com que são associadas a sinônimos positivos/negativos (e.g. [17]). A abordagem semântica também pode ser usada como complemento às outras abordagens, como forma de expansão ou de aquisição de vocabulário específico, na ausência de bons dicionários de sentimento. No entanto, ela carece ainda de métodos outros que manual, para validação da polaridade atribuída.

## 1.5. A Mineração de Opiniões e suas Fontes de Dados

### 1.5.1. Revisão de Produtos

Uma parcela significativa dos trabalhos na área de mineração de opiniões concentra-se na revisão de produtos, como já mencionado. Por um lado, este foco é justificado pelo interesse comercial nesta classe de aplicação, e a pela grande disponibilidade de dados. Mas do ponto de vista computacional, revisões de produtos apresentam uma série de propriedades que facilitam a mineração de opiniões, quando comparadas a textos em geral, tais como predominância de conteúdo de opinião, bom volume de opinião sobre uma mesma entidade ou relativa a um domínio, e definição clara da entidade alvo. Ainda, o uso de vocabulário muito coloquial, gírias, emoticons e mesmo de ironia é bem mais limitado, se comparado a outros tipos de mídia (e.g. Twitter, comentários). Assim, dado o foco claro, e a menor quantidade de ruído, os problemas computacionais inerentes podem ser mais facilmente identificados, estruturados e tratados [20]. Nesta seção, examinaremos trabalhos pioneiros nesta área [37, 24, 17], sendo que muitos outros resolvendo aspectos mais pontuais ou que melhoram resultados prévios são discutidos em [20, 34].

#### 1.5.1.1. *Analise de Opinião em Nível de Documento*

Um dos trabalhos precursores nesta área foi desenvolvido por Turney em [37]. A motivação deste trabalho é agregar ao resultado de uma busca sobre revisões de um produto/serviço, informação sobre a recomendação (*Thumbs up*) ou não (*Thumbs down*) deste. Suas principais características são: a) análise em nível de documento, b) uso de abordagem estatística para classificação de polaridade, c) identificação de porções de texto com sentimento usando informação morfológica.

No tocante à etapa de identificação, assume-se que uma mecanismo de busca retorna uma coleção dos documentos referentes à entidade consultada, a qual é considerada como alvo de todas opiniões expressas nos documentos. A etapa de classificação envolve, para cada documento, seu pré-processamento para identificação *features* de interesse com base nas classes morfológicas das palavras, e a definição de sua polaridade. Finalmente, esta proposta não envolve explicitamente a etapa de sumarização, mas os documentos polarizados podem ser a entrada para várias aplicações, tais como geração de estatísticas, identificação de fóruns com posts abusivos (*flames*), etc.

Um documento de entrada é analisado usando um *parser* de POS, que rotula cada

palavra do texto com sua classe morfológica. São então selecionados apenas pares de palavras que seguem determinados padrões sintáticos. Mais especificamente, retém-se pares de palavras onde um dos elementos é um adjetivo ou advérbio, e o outro, uma palavra que lhe dá contexto. Com o contexto, a polaridade de palavras de sentimento pode ser melhor avaliada, tais como o adjetivo "inesperado", que possui conotação positiva no contexto de filme ("final inesperado"), ou negativo em uma revisão sobre carros ("defeito inesperado").

Para cada expressão resultante do pré-processamento, é calculado o PMI-IR, com base no PMI da expressão analisada com as palavras *excelente* e *ruim*, respectivamente. Para cálculo do PMI são submetidas consultas à internet com a mecanismo de busca AltaVista, onde o número de documentos retornados é utilizado como cálculo de frequência. Esta disponibiliza o operador *near*, que restringe a busca a documentos onde os termos estão próximos dentro de uma janela de no máximo 10 palavras. A polaridade final do documento, denominada *orientação semântica*, é dada pela média da polaridade de todas as expressões consideradas.

Como avaliação, foram usadas 410 revisões extraídas do *site* Epinions sobre filmes, automóveis, bancos e destinos de viagem, as quais estão associadas com uma classificação de estrelas. Foi obtida uma acurácia média de 74,39% em relação à recomendação do produto/serviço avaliado. É interessante notar que os piores resultados obtidos foram no domínio de filmes, já que mesmo que o sentimento geral sobre o filme tenha sido bom, os usuários sempre criticaram negativamente aspectos específicos. Ainda, não foi possível distinguir a opinião sobre o filme, do seu conteúdo objetivo. Por exemplo, na revisão *"Ele falava de um jeito devagar e metódico. Eu amei! Isto fez ele mais arrogante e mais perverso."*, as expressões "mais arrogante" e "mais perverso" referem-se ao personagem, mas são interpretadas como opiniões sobre o filme. Essa situação não aconteceu nos outros contextos analisados.

O desempenho devido à grande quantidade de buscas necessárias, e os restritos padrões gramaticais considerados, são apontados como os fatores mais limitantes desta proposta.

Uma alternativa à etapa de classificação foi apresentada na mesma época por Pang et al. em [24], usando métodos de aprendizagem de máquina. Mais especificamente, o trabalho concentrou-se em comparar experimentos com três classificadores conhecidos (Support Vector Machine - SVM, Naïve Bayes e Maximum Entropy [32]), e diferentes alternativas de pré-processamento. Na preparação de um vetor de termos, foram consideradas as seguintes alternativas de pré-processamento, algumas delas combinadas:

- unigramas com representação binária: adoção de qualquer termo (após remoção das *stop words*, e sem *stemming*). Quando uma negativa aparecia próxima ao termo, foi adicionado um prefixo NOT_ para representar a negação. Os termos mais frequentes na coleção (aproximadamente 18.000) foram representados como um vetor binário de termos;

- unigramas com peso: semelhante aos unigramas acimas, mas com uma representação vetorial com pesos (TDF-IF);

- bigramas: foram selecionados os bigramas mais frequentes, sem negação, os quais foram representados como um vetor binário;
- uso de POS: os termos foram rotulados usando um *parser*. Experimentos foram feitos considerando somente adjetivos como *features*, ou prefixando cada termo com classe gramatical para dar mais contexto;
- uso de posição: uso de unigramas com indicação de sua posição no texto (início, meio e fim). A intuição é que pessoas iniciam o texto, ou o concluem com a opinião mais importante, e que portanto a posição das palavras de opinião pode ser importante.

Experimentos foram conduzidos sobre uma base de revisões sobre filmes, da qual foram extraídas 700 revisões positivas e 700 negativas. Em termo de qualidade, os resultados foram bastante semelhantes aos obtidos por Turney em [37]. Os experimentos não revelaram de forma consistente a superioridade de nenhum dos classificadores adotados, e em termos de preparação de dados, o pré-processamento de vetor binário de unigramas foi, de uma forma geral, o mais eficaz.

#### 1.5.1.2. *Análise de Opinião em Nível de Aspecto*

A análise de um produto em nível de documento permite derivar uma avaliação geral do mesmo. Em revisões onde existem sentimentos mistos expressos sobre a mesma entidade, pode-se chegar a uma situação de neutralidade em caso de médias (e.g. [37]), onde as expressões positivas e negativas se anulam; ou de incapacidade de classificação correta, pela falta de *features* discriminantes (e.g. [24]). A análise em nível de aspecto permite detalhar o alvo do sentimento, de tal forma que possam ser detectados seus pontos fortes e fracos.

Um trabalho pioneiro nesta área foi proposto por Hu e Liu em [17], que se caracteriza por: a) identificação de aspectos e de sentenças de opinião na etapa de identificação; b) uso de POS para identificação tanto de aspectos, quanto de palavras de sentimento; c) emprego da abordagem semântica para classificação da polaridade; e d) criação de sumários contendo o sentimento positivo e negativo sobre cada aspecto do produto.

A fase de identificação reconhece tópicos e sentenças de opinião. É primeiramente realizada a marcação das categorias gramaticais usando um *parser* de POS sobre os documentos. Para identificar os aspectos, são usadas regras de associação para encontrar termos frequentemente associados, os quais são podados segundo algumas regras de pré-processamento. Para encontrar as sentenças de opinião, são encontrados os adjetivos, considerados como palavras de opinião, e se estão próximos a aspectos, são designados como *opiniões efetivas*.

A etapa de classificação polariza as opiniões sobre os aspectos usando uma abordagem semântica, e refina o conjunto de tópicos utilizando as palavras de opinião. Para a polarização, o ponto de partida é um conjunto de palavras semente com polaridade definida manualmente. Usando o WordNet, a polaridade destas palavras é propagada para sinônimos (mesma polaridade) e antônimos (polaridade inversa), resultando assim em um

dicionário específico de sentimentos para o domínio das revisões. Este algoritmo de propagação é iterativo, de tal forma que a polaridade antes desconhecida de uma determinada palavra, acaba sendo encontrada em uma iteração futura, já que uma vez que um adjetivo é polarizado, ele passa a incorporar o conjunto de sementes. Para relacionar o sentimento ao aspecto, são usadas apenas as palavras de sentimento efetivas, na mesma sentença ou na sentença próxima.

As palavras de sentimento polarizadas são também utilizadas para identificar os aspectos de produtos menos frequentes e previamente desconhecidos. Por exemplo, se a palavra "excelente" é uma palavra de sentimento conhecida e na sentença, perto dessa palavra, existe uma frase nominal, assume-se que esta seja um aspecto de produto.

Finalmente, os sumários são criados em dois passos: 1) para cada aspecto, é feita uma contagem de quantas revisões opinam de maneira positiva/negativa; 2) aspectos são ordenados pela frequência com que aparecem nas revisões de produtos. Sentenças que contribuíram à pontuação positiva/negativa são mostradas junto com os sumários

A validação envolveu experimentos com vários aparelhos eletrônicos, usando revisões extraídas dos *sites* Amazon.com e Cnet.com. Estas foram manualmente anotadas quanto ao seus aspectos, sentenças opinativas e sua respectiva polaridade. Como resultado, obteve-se uma precisão média de 68,25% para aspectos identificados, de 64,2% para identificação de sentenças opinativas e uma acurácia média de 84,2% no tocante à polaridade. Os autores apontam como oportunidade de melhorias a resolução de pronomes, o uso de palavras de sentimento de classes outras que adjetivos, e o tratamento da intensidade da opinião sobre os aspectos extraídos.

### 1.5.2. Mineração de Opiniões em Notícias e Blogs

A mineração de opiniões em textos não estruturados é bem mais complexa que em revisões de produtos. Existem muitos desafios a serem enfrentados, entre eles: a) o texto pode conter opiniões sobre múltiplos alvos, e não é fácil reconhecê-los; b) o conteúdo de opinião é mais esparso no texto; c) dados os dois problemas anteriores, a associação da entidade alvo com a opinião fica ainda mais complexa; d) alguns tipos de texto, como notícias, tendem a não explicitar a opinião diretamente, fazendo-o através de artifícios (e.g. frases atribuídas a outras pessoas citadas na notícia); e e) é difícil distinguir entre conteúdo ruim (e.g. um terremoto), e uma opinião boa sobre um conteúdo ruim (e.g. elogiar o socorro às vítimas de um terremoto). Nesta seção são discutidos dois trabalhos que mineram opiniões sobre notícias: um que utiliza menções a entidades específicas [16], e outro que aborda os desafios de encontrar sentimento explícito em notícias [6, 7].

#### 1.5.2.1. *Monitoramento de entidades em jornais e blogs*

Godbole *et al.* em [16] têm como objetivo desenvolver um sistema de análise de sentimentos em notícias e blogs que monitore o sentimento do público geral em relação a determinadas entidades, como pessoas, locais ou marcas. Assume-se que as entidades analisadas possuem características singulares, tais como atletas, celebridades, políticos, criminosos, etc, e que as opiniões emitidas devem ser interpretadas dentro de cada contexto. O método propõe uma forma algorítmica de construir dicionários de sentimentos

voltados a cada contexto, e métricas para medir o sentimento expresso sobre essas entidades. As principais características desta proposta são: a) análise em nível de sentença; b) uso de menções específicas para identificação de entidades; c) uso de abordagem semântica de criação de léxicos voltados a cada domínio, os quais são usados na polarização; e d) representação do sentimento através de diferentes métricas, cuja evolução pode ser monitorada.

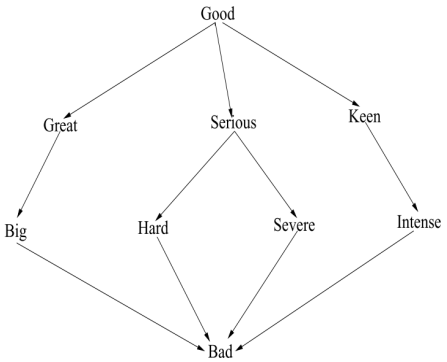


**Figura 1.6. Expansão da palavra "*good*" até uma inversão de sentimento (Fonte: [16]).**

Na fase de identificação, são encontradas as menções às entidades de interesse, e as respectivas sentenças. O método de co-ocorrência é utilizado para encontrar variações nas menções à mesma entidade, e uma ferramenta é utilizada para resolver pronomes. Usando os léxicos específicos relativos ao domínio da entidade em questão, cada sentença é polarizada. Se a entidade e uma palavra do léxico estiverem na mesma sentença, sua polaridade é atribuído ao alvo.

Para a fase de sumarização, são propostas métricas de sentimento, envolvendo *polaridade* e *subjetividade*. A primeira busca determinar a razão entre sentimento positivo e negativo, enquanto que a segunda, a razão entre sentimento positivo e o total de sentimento (positivo e negativo). Com isto, podem ser medidas a polaridade e subjetividade relativa a uma dada entidade, ou em relação ao conjunto de todas entidades (denominado mundo). A intuição é que determinadas épocas (e.g. natal, eleições) implicam que mais sentimentos sejam expressos. A evolução destas métricas pode ser acompanhada visualmente em um gráfico.

Os autores ainda fazem uma comparação entre o conteúdo de blogs e jornais para as mesmas entidades, concluindo que o sentimento relacionado a certas entidades pode diferenciar-se em notícias e blogs, devido ao viés do meio de publicação.

O método foi validado utilizando indicadores externos existentes (e.g. sentimento detectado sobre o presidente, e pesquisas de intenção de votos) mostrando que os resultados são bastante próximos.

### 1.5.2.2. *Análise de sentimentos em notícias*

O foco do trabalho de Balahur *et al.* [6, 7] é a identificação de conteúdo subjetivo em notícias. Exceto por jornais sensacionalistas, este tipo de texto evita expressar explicitamente

opiniões, com objetivo de manter a seriedade e a suposta neutralidade da notícia. Opiniões neste meio são expressas de formas bem mais sutis, e que são linguisticamente difíceis de serem identificadas: argumentações tendenciosas; omissão ou destaque de fatos em detrimento de outros; citações de outras pessoas que expressam opiniões (e.g. "reiniciar este julgamento é um verdadeiro absurdo", disse o Ministro da Justiça); entre outros.

Dada a complexidade da análise de conteúdo implícito, a análise da subjetividade restringe-se a opiniões explicitamente expressas em citações, nas quais o conteúdo potencialmente de opinião é associado a um autor. Balahur *et al.* enfatizam a complexidade da mineração de opinião neste tipo de texto, e a restrição a citações como um primeiro passo para a compreensão dos diferentes problemas envolvidos. Este trabalho caracteriza-se por: a) nível de análise em citação, composta de uma ou mais sentenças; b) abordagem orientada a léxicos; c) uso de menções específicas para identificação de entidades.

A principal contribuição destes trabalhos é uma melhor caracterização da tarefa de mineração de opiniões em citações, através da realização de experimentos. As tarefas são definidas como:

- a definição do alvo da opinião: mesmo restringindo a citações, o alvo da opinião nem sempre é claro e explícito, podendo existir mais de um alvo. Utilizam-se entidades específicas como alvo, que estão em uma dada janela de distância das citações. As entidades são termos de busca de notícias (e.g. tsunami, Obama);

- a separação entre conteúdo ruim/bom, da opinião positiva ou negativa sobre este: a abordagem é baseada em léxicos, dos quais são removidos termos específicos que podem representar fatos com conotação negativa/positiva (e.g. crise, desastre, carnaval). Nos experimentos, foram utilizados rótulos (*tags*) que caracterizam categorias de notícias, e que possuem um sentimento associado. Os autores apontam que outras abordagens poderiam ter sido utilizadas com melhores resultados, inclusive escolha manual de termos.

- interpretação da polaridade sem informação contextual: os autores discorrem que em notícias existe uma grande distinção entre a opinião do autor, expressa na citação, e a opinião do leitor, de acordo com sua interpretação dos fatos. Esta divergência é uma grande dificuldade para o consenso em processos de anotação. Eles focam na classificação do primeiro tipo, usando apenas as informações explicitamente presentes no texto. Para a polarização, que foi baseada na abordagem de dicionário, foram experimentados 4 léxicos distintos, utilizados para polarizar todos os termos de sentimento encontrados na citação a uma dada distância do alvo. Diferentes abrangências de janela foram testadas.

Os experimentos envolveram 1.292 citações sobre as quais houve concordância entre os anotadores sobre a polaridade e o alvo. O melhor resultado (86% de acurácia) foi obtido com o uso combinado de dois léxicos (entre eles JRCTonality, desenvolvido pelo próprio grupo) e uma janela de 6 palavras entre alvo e termo de sentimento. Na análise de erros, os principais problemas detectados foram as citações polarizadas erroneamente como neutras (devido à falta de palavras explícitas), o uso de ironias, emprego de abstrações para referir-se ao alvo, e múltiplos alvos.

### 1.5.3. Mídia Social

A mineração em mídias sociais têm a informalidade deste tipo de meio como um dos seus principais desafios. Não apenas o vocabulário pode ser bem específico e volátil, como o número de erros de digitação, de ortografia ou gramaticais pode invalidar a contribuição de análises linguísticas. Por outro lado, o volume da dados gerados sobre cada tópico é tão grande, quando comparado com outras fontes, que tais erros podem não ser relevantes. Um dos focos de trabalhos de mineração de opinião no Twitter, é o da capacidade de previsão, justamente com base neste grande volume. Dois destes trabalhos [3] [8] são melhor detalhados no restante desta seção.

#### 1.5.3.1. Previsão de indicadores de rentabilidade de filmes

A popularidade do Twitter é tal que organizações têm utilizado esta plataforma para divulgação e marketing de suas marcas e produtos. Este é o caso de filmes, onde produtores têm investido massivamente em publicidade e marketing voltado aos usuários do Twitter. Asur *et al.* em [3] realizam experimentos para verificar se o interesse que um filme desperta, particularmente na época de pré-lançamento, correlaciona-se com indicadores econômicos deste domínio. Mais especificamente, a partir do volume de *tweets* e do sentimento neles expressos, deseja-se prever: a arrecadação da primeira semana de bilheteria; os preços dos índices do *Hollywod Stock Exchange* (HSX); e o rendimento de todos os filmes de uma semana específica. Este trabalho caracteriza-se por: a) análise em nível de documento (*tweet*); b) uso de termos pré-definidos para determinação de entidades alvo; c) uso da abordagem de aprendizagem de máquina para polarização, e d) definição de métricas de agregação de sentimento utilizadas para previsão.

A proposta consiste de um estudo de caso onde foram analisados *tweets* de filmes específicos, sobre um período de três meses. Palavras-chaves, como URL's de material promocional e o título dos filmes, foram utilizadas para extrair os *tweets* relevantes e identificar os alvos. Juntamente com o conteúdo de cada *tweet*, também extraiu-se a data e hora de postagem, e o respectivo autor.

Os autores comparam dois tipos de previsão: quantitativo (quantidade de *tweets* e *retweets* contendo URL's de material promocional sobre o filme), e baseado em opinião expressa nos *tweets*. A segunda implica a necessidade de classificação da polaridade dos *tweets*, que foi realizada utilizando o classificador DynamicLMClassifier, disponível no pacote análise linguístico LingPipe[6]. Este classificador é ternário, i.e. classifica os *tweets* como positivo, negativo ou neutro. Os dados de treino foram rotulados manualmente utilizando-se o Amazon Mechanical Turk. Somente *tweets* com polaridade classificada de forma unânime pelos anotadores foram utilizados para treino. A acurácia do classificador foi elevada (98%) com features representando 8-gramas.

Para a previsão baseada em número de *tweets*, os autores definem a métrica taxa de *tweets* de cada filme por hora (Equação 4). Esse métrica permite criar uma série temporal de menções de cada filme, e correlacioná-la usando regressão linear com as variáveis alvo, ou seja, previsão de bilheteria e previsão de valor dos índices HSX. Uma forte correlação

[6]http://www.alias-i.com/lingpipe

foi encontrada ($R^2$ ajustado de 0,8, e de 0,9, respectivamente), significando alto poder de previsão.

$$\textit{Taxa-de-tweets(filme)} = \frac{|\textit{tweets(filme)}|}{|\textit{tempo (em horas)}|} \quad (4)$$

Já para a previsão baseada em sentimento, foram definidas duas outras métricas: a razão entre o total de *tweets* positivos e negativos e o total de *tweets* neutros, chamada pelos autores de *subjetividade*; e a razão entre *tweets* positivos e negativos. Os resultados obtidos foram inferiores aos obtidos com a métrica quantitativa de taxa de *tweets* por hora. No entanto, quando as métricas de sentimento são associadas à da taxa de *tweets*, houve uma melhoria no poder de predição.

É interessante notar que resultados semelhantes foram encontrados no domínio político, onde experimentos revelaram que a quantidade de *tweets* tiveram maior poder preditivo sobre o resultado de eleições, que o sentimento ou emoção neles expressos [21, 35].

#### 1.5.3.2. Previsão do comportamento da bolsa de valores

O objetivo deste trabalho é semelhante ao da seção anterior, no domínio da bolsa de valores. Bollen *et al.* [8] realizam experimentos para verificar se sentimento expresso no Twitter, chamado no trabalho de humor, tem influência sobre a bolsa de valores, e pode ser utilizado para prever seu comportamento usando o índice Dow Jones. Este trabalho caracteriza-se por: a) análise em nível de documento (*tweet*); b) uso de expressões específicas para determinar conteúdo de sentimento; c) uso da abordagem de léxica para polarização, onde um léxico de emoções também foi usado, e d) definição de métricas de agregação de sentimento utilizadas para previsão.

Foi realizado um estudo de caso envolvendo *tweets* correspondentes a um período de onze meses (Fevereiro, 2008 - Dezembro, 2008). A fase de identificação selecionou apenas *tweets* que continham sentimento explícito. Como critério, foram utilizadas expressões pré-definidas como “eu sinto”, “eu estou sentindo”, “eu não sinto”, etc.

A classificação envolveu a abordagem léxica, mas dois tipos de classificação foram feitos: a) polaridade de sentimento (positivo e negativo), utilizando o léxico OpinionFinder [38]; e b) classificação da emoção (vide Seção 1.2.1), utilizada uma ferramenta denominada Google-Profile of Mood States (GPOMS). GPOMS analisa e classifica a emoção em seis diferentes dimensões: *calma, alerta, certeza, vitalidade, gentileza* e *felicidade*.

A polaridade da opinião foi sumarizada através de uma métrica que representa a razão entre a quantidade de termos positivos encontrados nos *tweets* e a de termos negativos, em um determinado dia. Já cada dimensão da emoção GPMOS foi totalizada por dia. Para todas estas métricas, foram criadas séries temporais, as quais foram comparadas com eventos conhecidos do mesmo período: eleições e Ação de Graças. Exceto pela *gentileza* e *alerta* em relação às eleições, foi demonstrado que as técnicas utilizadas caracterizavam o humor típico destas datas. Uma análise estatística mostrou ainda correlação entre a po-

laridade da opinião e as dimensões de emoção *certeza*, *vitalidade*, e *alegria*, mas não com as demais.

Finalmente, foi usado um método baseado em redes neurais fuzzy (*self-organizing fuzzy neural network* - SOFNN) para correlacionar as séries temporais das polaridades das opiniões e das emoções, com a série temporal Down Jones Industrial Average (DIJA), e desenvolver um modelo preditivo. A abordagem utilizando o OpinionFinder não obteve bons resultados. No entanto, as emoções *calma* e *felicidade* demonstraram correlação com o DIJA em certos trechos do período analisado. O sentimento de calma foi o que obteve um melhor poder preditivo com um atraso (*lag*) de 3 dias. No entanto, o sentimento de calma só conseguiu prever corretamente o DIJA em períodos em que não há eventos inesperados (e.g. anúncio da Reserva Federal Americana). Isto significa dizer que consegue prever quando os índices se mantêm devido à falta de eventos externos importantes. Os autores concluem que a polaridade da opinião pode ser muito abrangente, e esconder aspectos cobertos pela subjetividade das emoções.

## 1.6. Conclusões e Direções Futuras

A mineração de opiniões é uma área de crescente interesse. Neste capítulo, discutimos seus conceitos básicos, os desafios na detecção de sentimento e de seu alvo, e técnicas que podem ser usadas para identificar, classificar a polaridade e agregar o sentimento expresso. Alguns trabalhos clássicos envolvendo diferentes fontes de opiniões foram apresentados a título de ilustração. O volume crescente de conteúdo subjetivo disponível diariamente, em particular nas redes sociais, motiva o crescimento da área com novas técnicas capazes de processar automaticamente textos, de forma escalável, robusta, precisa e independente de domínio e de linguagem. Muitas são as aplicações centradas na sumarização e visualização do sentimento, ou na predição de comportamentos com base no sentimento existente. Empresas, eventos (e.g. Olimpíadas), personalidades, estão interessadas na compreensão de como são percebidas pelo público em geral em tempo real, e nas mais variadas mídias.

A área ainda apresenta muitos problemas e oportunidades. Muitos esforços estão voltados à detecção de outros tipos de opinião: comparativa, implícita, dependente do observador, contraditórias, spams, ironias e sarcasmos, etc. A identificação de alvos e opiniões em mídias sem grau de estruturação, como notícias, é também outra área bastante importante. O desenvolvimento de recursos multilíngues permitirão o avanço do estado da arte, permitindo tratar corpus para os quais hoje não existem recursos (e.g. dados rotulados, léxicos, recursos para tratamento da língua natural, etc.). A evolução das técnicas de classificação para métodos escaláveis, menos sensíveis ao contexto ou a ruídos, e que combinam abordagens já existentes em um *framework* único é uma outra meta. Novas aplicações devem estabelecer soluções para *streaming* de dados, apoio a decisão baseado em sentimento, predições, entre tantas outras.

## Referências

[1] Aggarwal, C. C. and Zhai, C. (2012). *Mining text data*. Springer.

[2] Archak, N., Ghose, A., and Ipeirotis, P. G. (2007). Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proceedings*

*of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 56–65. ACM.

[3] Asur, S. and Huberman, B. A. (2010). Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE.

[4] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204.

[5] Balahur, A., Kozareva, Z., and Montoyo, A. (2009a). Determining the polarity and source of opinions expressed in political debates. In *Computational Linguistics and Intelligent Text Processing*, pages 468–480. Springer.

[6] Balahur, A., Steinberger, R., Goot, E. v. d., Pouliquen, B., and Kabadjov, M. (2009b). Opinion mining on newspaper quotations. In *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, volume 3, pages 523–526. IET.

[7] Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Van Der Goot, E., Halkia, M., Pouliquen, B., and Belyaeva, J. (2010). Sentiment analysis in the news. In *Proceedings of LREC*, volume 10.

[8] Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

[9] Bruce, R. F. and Wiebe, J. M. (1999). Recognizing subjectivity: a case study in manual tagging. *Natural Language Engineering*, 5(2):187–205.

[10] Calais Guerra, P. H., Veloso, A., Meira Jr, W., and Almeida, V. (2011). From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158. ACM.

[11] Carvalho, P., Sarmento, L., Silva, M. J., and de Oliveira, E. (2009). Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

[12] Dave, K., Lawrence, S., and Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM.

[13] Dey, L. and Haque, S. (2009). Opinion mining from noisy text data. *International journal on document analysis and recognition*, 12(3):205–226.

[14] Fellbaum, C. (2010). *WordNet*. Springer.

[15] Ghani, R., Probst, K., Liu, Y., Krema, M., and Fano, A. (2006). Text mining for product attribute extraction. *ACM SIGKDD Explorations Newsletter*, 8(1):41–48.

[16] Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, volume 2.

[17] Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM.

[18] Ku, L., Liang, Y., and Chen, H. (2006). Opinion extraction, summarization and tracking in news and blog corpora. In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, number 2001.

[19] Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:568.

[20] Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers.

[21] O'Connor, B., Balasubramanyan, R., Routledge, B. R., and Smith, N. A. (2010). From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129.

[22] Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*.

[23] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

[24] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

[25] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., and Booth, R. J. (2007). The development and psychometric properties of liwc2007. *Austin, TX, LIWC. Net*.

[26] Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases*, 1(3):261–377.

[27] Sarmento, L., Carvalho, P., Silva, M., and de Oliveira, E. (2009). Automatic creation of a reference corpus for political opinion mining in user-generated content. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 29–36. ACM.

[28] Silva, M., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. *Computational Processing of the Portuguese Language*, pages 218–228.

[29] Souza, M., Vieira, R., Busetti, D., Chishman, R., and Alves, I. M. (2011). Construction of a portuguese opinion lexicon from multiple resources. In *The 8th Brazilian Symposium in Information and Human Language Technology (STIL 2011), Cuiabá, Brazil*.

[30] Stone, P. J., Dunphy, D. C., and Smith, M. S. (1966). The general inquirer: A computer approach to content analysis.

[31] Strapparava, C. and Valitutti, A. (2004). Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086.

[32] Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to data mining*. Addison Wesley.

[33] Thet, T., Na, J., and Khoo, C. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science*, 36(6):823–848.

[34] Tsytsarau, M. and Palpanas, T. (2012). Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514.

[35] Tumasjan, A., Sprenger, T., Sandner, P., and Welpe, I. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International aaai conference on weblogs and social media*, pages 178–185.

[36] Tumitan, D. and Becker, K. (2013). Tracking sentiment evolution on user-generated content: A case study in the brazilian political scene. In *Brazilian Symposium on Databases (SBBD)*, page 6. SBC. A ser publicado.

[37] Turney, P. D. (2002). Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 417–424. Association for Computational Linguistics.

[38] Wiebe, J. and Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Computational Linguistics and Intelligent Text Processing*, pages 486–497. Springer.

[39] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.