PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

MINERAÇÃO DE OPINIÕES APLICADA A MÍDIAS SOCIAIS

MARLO VIEIRA DOS SANTOS E SOUZA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Renata Vieira

Porto Alegre
2012

S729m Souza, Marlo Vieira dos Santos e
Mineração de opiniões aplicada a mídias sociais / Marlo Vieira dos Santos e Souza. – Porto Alegre, 2012.
76 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Profª. Drª. Renata Vieira.

1. Informática. 2. Processamento da Linguagem Natural. 3. Recuperação da Informação. I. Vieira, Renata. II. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada "***Mineração de Opiniões Aplicada a Mídias Sociais***", apresentada por Marlo Vieira dos Santos e Souza como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 19/03/2012 pela Comissão Examinadora:

Profa. Dra. Renata Vieira - PPGCC/PUCRS
Orientadora

Profa. Dra. Vera Lúcia Strube de Lima - PPGCC/PUCRS

Profa. Dra. Renata Galante - UFRGS

Homologada em...../...../....., conforme Ata No. ..... pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

"Lines slip easily down the accustomed grooves. The old designs are copied so glibly that we are half inclined to think them original, save for that very glibness. "

Virgínia Woolf, A Letter to a Young Poet

AGRADECIMENTOS

Primeiramente, à minha família pelo amor e apoio em todos os momentos.

Aos amigos e amigas, daqui e de lá, pela paciência, pelos risos e pelas muitas horas de telefonemas.

À minha orientadora Renata, quase uma segunda (ou seria terceira?) mãe, que muito me ensinou e a quem eu aprendi a respeitar e escutar. Obrigado por brigar comigo, por estar sempre lá quando eu precisava e por sempre me escutar, mesmo quando você não concordava.

Aos colegas do laboratório PLN, Daniela, Lucas, Rodrigo, Paulo, Fernando, Igor, Roger, Lucelene, Daniel e Guilherme, em especial, Clarissa, Sandra, Larissa, Patrícia, Mírian e Douglas pelas dicas, pelo companheirismo e pelas boas doses de café.

À professora Vera do PPGCC da PUCRS por ter contribuído tanto para a realização deste trabalho, desde sua gênese. Sem sua crítica aguçada, creio que o trabalho não teria sido tão proveitoso.

À Plugar e à Prognus, pelo apoio financeiro durante o mestrado.

"E ó: prum monte de gente aí risos #bgogalero "(sic)

MINERAÇÃO DE OPINIÕES APLICADA A MÍDIAS SOCIAIS

RESUMO

O ambiente competitivo se tornou, nas ultimas décadas, mais dinâmico graças às tecnologias de informação e comunicação e à globalização. O gestor, assim, precisa estar sempre bem informado sobre o panorama competitivo antes de tomar decisões estratégicas. Nessa direção, a Inteligência Competitiva (IC) surge como uma disciplina que pretende sistematizar a obtenção e análise de informações do ambiente competitivo com função de auxiliar a tomada de decisão.

Há entretanto uma quantidade crescente de informação sendo produzida e disponibilizada em meios como a Internet e mídias tradicionais, as quais se tornam de difícil manejo. Associado a isso, os gestores sofrem ainda com restrições temporais para responder ao estímulo do mercado e manterem-se competitivos. Dessa forma, é necessário manter uma equipe de monitoramento constante do ambiente competitivo para que se possa lidar com a quantidade de informação proveniente de diversas fontes. Acreditamos que a aplicação de técnicas de Análise de Texto podem auxiliar nas diversas fases do processo de IC.

O presente trabalho apresenta uma proposta de utilização de tais técnicas para auxiliar o processo de Inteligência Competitiva. Discutimos aqui a utilização de um método de Análise de Sentimentos aliado ao Reconhecimento de Entidades Nomeadas em textos provenientes de mídias sociais - particularmente o Twitter - que permitam analisar as atitudes do mercado consumidor quanto a uma determinada marca.

São apresentados ainda o sistema desenvolvido, as avaliações realizadas e as conclusões que tiramos.

Palavras-chave: Análise de Sentimentos; Entidades Nomeadas; Twitter; Língua Portuguesa.

OPINION MINING IN SOCIAL MEDIA

ABSTRACT

The competitive environment has become more dynamic in the last few decades due to the great development of information and comunication technologies and to the globalization process.A company manager must, thus, always be well informed about the competitive landscape before making strategic decisions. In this sense, the Competitive Intelligence (CI) emerges as a discipline that aims to systematize the collection and analysis of information in the competitive environment willing to assist decision making.

There is, however, an increasing amount of information being produced and released in Internet and traditional media, which become unwieldy. Associated with this, managers still suffer with time constraints to respond to the market stimuli and remain competitive. Thus, it is necessary to maintain a constant staff monitoring the competitive environment to be able to handle the amount of information from this various sources. We believe that the application Text Analysis techniques can help in various stages of such process.

This work presents a proposal to use such techniques to aid the process of Competitive Intelligence. We discuss the use of Sentiment Analysis techniques coupled with Named Entity Recognition in texts from social media - especially Twitter - which helps in the analysis of the attitudes of the consumer market towards a brand.

We also present a system implementing the proposed techniques, the evaluations made with it and present our conclusions.

Keywords: Sentiment Analysis; Named Entities; Twitter; Portuguese Language.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

ACE	*Automatic Content Extraction*
AS	Análise de Sentimentos
BILOU	Begin, Inside, Last, Outside, Unit
BIO	Begin, Inside, Outside
HAREM	uma Avaliação conjunta para o Reconhecimento de Entidades Mencionadas
IC	Inteligência Competitiva
KNN	*K nearest neighbors*
MUC	*Message Understanding Conference*
NLTK	*Natural Language Toolkit*
PLN	Processamento de linguagem natural
POS	*Part-of-Speech*
REN	Reconhecimento de Entidades Nomeadas
SVM	*Support Vector Machines*
WOM	*Word of Mouth*

SUMÁRIO

1. Introdução

Neste capítulo apresentamos a motivação, o contexto, o objetivo e a organização deste trabalho.

1.1 Motivação e Contexto do Trabalho

Levy [46] argumenta que, após a Segunda Guerra Mundial, a informação e o conhecimento tornaram-se a principal fonte de produção de riquezas, tornando-se os bens econômicos primordiais. Nesse contexto, na sociedade em rede, pode-se entender a atividade econômica como um processo de Inteligência Coletiva, onde mesmo o consumo é atividade produtora, construindo informação.

Aquele autor, em [47], estabelece como uma característica fundamental da Inteligência Coletiva a passagem de uma mídia de massa na qual poucos têm a capacidade de produzir informação, enquanto muitos atuam como meros consumidores, para um paradigma em que todos participam consumindo e produzindo informação. A interconexão em rede potencializa a troca incessante e bilateral de informações, transformando o receptor passivo em produtor de conteúdo, caracterizando a criação dos *new media* (novas mídias, em português), em contraste ao *mass media* (mídias de massa, ou mídias massivas em português).

A esse novo tipo de mídia, em que muitos produzem para muitos, chama-se Mídia Social. As mídias sociais apresentam-se de várias formas: listas de discussão, blogosfera, sites de redes sociais (como Facebook[1], Twitter[2], Orkut[3], etc.), entre outros.

Dourado [22] considera que as mídias sociais representam uma transformação na mediação da informação ao eliminar o papel do intermediário no esquema comunicacional das empresas com seus clientes. A comunicação direta com os clientes permite saber o perfil de seu consumidor, assim como ter acesso à sua opinião a fim de oferecer os produtos e serviços por eles demandados.

Diante de mercados extremamente dinâmicos e complexos, devido ao desenvolvimento tecnológico e a globalização, tomar decisões estratégicas em uma empresa tornou-se uma atividade cercada de incertezas. Para garantir que uma decisão seja a mais apropriada, dado o panorama competitivo, o gestor deve estar muito bem informado sobre o ambiente interno e externo da companhia. De fato, Porter [69] defende que, para definir a estratégia competitiva da empresa, um ponto fundamental é a análise detalhada do ambiente competitivo em que a empresa se situa.

Para formalizar as práticas de monitoramento e análise do ambiente competitivo surge a disciplina da Inteligência Competitiva. Dada esta necessidade de se analisar o ambiente competitivo, a fim de se tomar decisões estratégicas e operacionais, e a centralidade do papel da marca entre os ativos da empresa para se alcançar uma vantagem competitiva sustentável [1], o monitoramento de marcas é uma importante tarefa a ser realizada dentro desse processo. A informação disponível nas diversas mídias é então de extrema importância para uma organização situar-se no panorama competitivo.

[1] http://www.facebook.com
[2] http://www.twitter.com
[3] http://www.orkut.com

Toda essa informação produzida e disponibilizada em meios como a Internet e mídias tradicionais são, entretanto, de difícil manejo. A especial dificuldade de tratá-las deriva do fato que as mesmas se encontram em formas não estruturadas ou semi-estruturadas, como em texto e páginas de Internet.

Uma vez que a quantidade de informação disponível sobre o mercado cresceu sobremaneira pelos novos recursos de indexação e distribuição de informações, é necessário utilizar-se técnicas de acesso e manipulação desses dados e informações, a fim de se poder analisá-los de forma consistente. Devido à natureza não estruturada dos dados, o tratamento computacional desses é de difícil realização. Nesse contexto, o papel das tecnologias de informação e, em particular, da Análise de Texto (do inglês *Text Analytics*) , área que procura extrair informação a partir de coleções textuais, são de fundamental importância para auxiliar tal processo.

Nossa convicção é que a aplicação de técnicas de Análise de Texto pode auxiliar o processo de Inteligência Competitiva. Acreditamos que utilizando a Análise de Sentimentos - o processo de identificar (ou extrair) emoções, opiniões ou pontos de vista automaticamente - podemos auxiliar o processo de monitoramento da marca, ao se minerar as Mídias Sociais - em especial ao Twitter -, desde que tal análise seja feita em um nível adequado. Esse monitoramento já é, de fato, realizado em diversas empresas, entretanto é ainda feito de forma manual ou semi-automatizada. Não encontramos uma ferramenta que realize a tarefa pretendida numa profundidade que consideramos adequada para a análise, i.e. a identificação de opinião a um nível de entidade.

Propomos com esse trabalho não somente identificar uma opinião expressa em texto, mas também relacioná-la com a entidade à qual ela se direciona. Escolhemos tal nível de análise devido à centralidade do papel da marca no panorama competitivo a ser analisado e da precariedade das ferramentas atualmente disponíveis para identificar a correta referência de uma opinião. A marca representa, como Aaker [1] salienta, o principal ativo da empresa e precisa de grande manutenção. Posta essa necessidade, acreditamos que a correta identificação da referência de uma opinião tem particular importância nesse panorama, uma vez que expõe diretamente a atitude do consumidor quanto à marca da corporação - ou de suas concorrentes.

1.2 Objetivo do Trabalho

Aplicar um método de análise de sentimentos focado em entidades para monitoramento de marcas ao minerar uma base de textos de uma mídia social - o Twitter - e avaliá-lo a partir da implementação de um sistema para a língua portuguesa.

1.3 Organização da Dissertação

O restante do trabalho está organizado como segue: no Capítulo 2, a fundamentação teórica é apresentada; no Capítulo 3, discutimos os trabalhos relacionados ao nosso e as semelhanças e diferenças entre eles. No Capítulo 4, apresentamos nossa estratégia para a resolução do problema, no Capítulo 5 são apresentados os resultados obtidos e, por fim, no Capítulo 6 apresentamos nossas considerações finais.

2. Fundamentação Teórica

Dada a importância do tema e apesentada a motivação do nosso trabalho, passamos então a apresentar a base teórica que utilizamos para entender (e atacar) o problema proposto. Neste capítulo,então, introduzimos a fundamentação teórica do nosso trabalho explicitando os principais conceitos e técnicas utilizadas posteriormente.

2.1 Inteligência Competitiva e monitoramento de marcas em Mídias Sociais

A Inteligência Competitiva (IC) é definida em [66] como o processo de monitorar o ambiente competitivo. Ela permite a profissionais tomarem decisões embasadas em informações atualizadas do ambiente externo e interno da organização. Para Oliveira [61], o objetivo da IC é prover informações estratégicas para reduzir a incerteza associada a uma decisão. Gomes e Braga [30] a definem como a permanente avaliação do ambiente competitivo e dos recursos de que se dispõe. Para esses autores, os objetivos da IC são antecipar mudanças ambientais, descobrir potenciais novos concorrentes, antecipar ações da concorrência e auxiliar aquisições e fusões. Fuld [26], por sua vez, define a inteligência competitiva como informação analisada que fornece *insights* e vantagem.

Pode-se então definir a Inteligência Competitiva como uma estratégia de monitorar o ambiente competitivo de uma determinada organização com finalidade de auxiliar a tomada de decisão estratégica ou tática. Ela tentar formalizar o processo de aquisição, análise e disseminação do conhecimento sobre o mercado em que a organização está inserida, possibilitando, aos gestores, a tomada de decisão informada.

O processo de inteligência é um processo sistemático e formalizado de se obter informações e analisá-las para gerar inteligência competitiva. Gomes e Braga [30] estabelecem um modelo de Inteligência Competitiva com cinco passos, baseando-se no modelo tradicional de IC de Herring [35]:

- Identificação das necessidades de informação: etapa em que se definem as necessidades ou requisitos de informação necessários, assim como as questões estratégicas que devem ser respondidas;

- Coleta das informações: fase em que se identificam e classificam as fontes de informação e se realiza a coleta, propriamente dita. Cada fonte de informação é classificada em relação à natureza, à origem e, principalmente, à confiabilidade;

- Análise das informações: etapa em que "o analista[, munido de modelos de análise,] transforma as informações coletadas em avaliação significativa, completa e confiável" [30, p.61]. O objetivo da análise é responder às questões estratégicas levantadas na primeira fase do processo de inteligência;

- Disseminação: envolve a entrega dos resultados da análise aos tomadores de decisão;

- Avaliação: fase de auto-análise do processo de inteligência, na qual se reflete sobre as fases realizadas e seu andamento, assim como sobre a usabilidade dos resultados para a organização.

Este trabalho propõe a aplicação de técnicas de Processamento de Linguagem Natural (PLN) para a fase de análise, com foco em identificação de opinião para entidades nomeadas. Dito isso, deixamos claro que o planejamento e coleta devem ter sido realizados e, portanto, sabe-se o que analisar e possui-se os dados para tal análise. Na fase de análise, os dados e informações serão organizados, estruturados e analisados para permitir a geração de *insights* competitivos e relações entre os dados até então não conhecidas, que ajudarão posteriormente no processo de tomada de decisão.

2.1.1 Mídias Sociais

Mídias sociais são as formas de nova mídia caracterizadas pelo *Socialcast*, i.e. o paradigma em que muitos transmitem informação para muitos, em oposição às mídias tradicionais, ou de massa, em que um transmite para muitos. Kaplan e Haenlein [39] as definem como "um grupo de aplicações baseadas na Internet construídas sobre as fundações ideológicas e técnicas da Web 2.0, que permite a criação e troca de conteúdo gerado pelo usuário" (tradução nossa).

Essa característica das mídias sociais - a saber, o conteúdo ser produzido pelos usuários - é justamente o que permite explorar essas novas mídias e configurar um novo espaço de comunicação para uma empresa ou organização, não mais centrada em falar para seu mercado, mas numa interação dialógica, onde se pode escutar diretamente esse mercado. Isso as torna um instrumento potencializador de Inteligência Coletiva pois permitem a interação social dos indivíduos em rede, trocando informação de forma dialógica e instantânea, não limitados pelas restrições geográficas. Elas apresentam-se de várias formas: listas de discussão, blogosfera, sites de redes de sociais (como Facebook[1], Twitter[2], Orkut[3], etc.), entre outros.

Dourado [22] considera que as mídias sociais representam uma transformação na mediação da informação ao eliminar o papel do intermediário no esquema comunicacional das empresas com seus clientes. A comunicação direta com os clientes permite saber o perfil de seu consumidor, assim como ter acesso à sua opinião a fim de oferecer os produtos e serviços por eles demandados. Nesse trabalho a autora discute novos modelos de negócio focando os *Social Media*.

Para Kaplan e Haenlein [39], a entrada nesse espaço por parte das empresas envolve novas formas de pensar, mas possui grande eficiência a um baixo custo para o contato com o usuário final, sendo assim importante para empresas grandes e pequenas. Os autores dão um conjunto de conselhos para as empresas que desejem utilizar mídias sociais.

Para Pompéia [67], a inserção de uma marca nas mídias sociais não está sob o controle da organização a qual pertence, mas, antes, depende de seus clientes. A autora argumenta que na política do *Socialcasting*, que caracteriza as mídias sociais, é importante para as empresas procurar

[1] http://www.facebook.com
[2] http://www.twitter.com
[3] http://www.orkut.com

relacionamento, interação com seus clientes, não audiência. Fica clara a necessidade de planejamento e de análise da imagem da empresa e da marca para criar seu plano de comunicação.

É patente o *locus* privilegiado que as mídias sociais têm na vida das pessoas, em particular no contexto brasileiro, onde informações que tornam-se populares nesse tipo de mídia chegaram a migrar para as ditas mídias tradicionais[4]. Por esse motivo entender o comportamento dos usuários dessas mídias, i.e. entender o próprio funcionamento da mídia e como ela é significada pelos usuários, é um importante passo para uma organização poder inserir-se na mesma. Nesse contexto, os estudos sobre a adoção de diversas mídias sociais no contexto do usuário brasileiro tornam-se de especial importância e são tratados a seguir.

2.1.2 Mídias Sociais no Brasil

Em diferentes estudos, o comportamento do usuário brasileiro foi mapeado em diferentes plataformas da Internet, como a blogosfera [71, 73, 74, 78], Orkut[5] [72, 73] e Twitter[6] [75–77]. Uma importante descoberta desses estudos diz respeito à apropriação que o usuário brasileiro faz dessas ferramentas na manutenção da sua rede social e construção do que os autores chamam de capital social - o capital associado ao pertencimento de uma coletividade.

É de se notar que as plataformas sociais na Web possuem características distintas e formas de apropriação distintas. Algumas possuem um perfil mais informacional, no sentido que os usuários se apropriam desta ferramenta no intuito de produzir e transmitir informação aos outros, enquanto outras possuem um perfil mais relacional, onde o foco do usuário é estabelecer relações entre si.

Ferramentas como o Orkut, por exemplo, possuem um perfil mais relacional, onde os usuários procuram adquirir capital social relacional e cognitivo, ou seja capital social relacionado às ligações sociais da rede [73].

Tais características no perfil de interação de uma determinada rede são importantes pois auxiliam a decidir qual tipo de inserção uma determinada organização deve ter na mesma e quais informações podem ser mineradas para conseguir vantagem competitiva. Uma mídia altamente informacional, com bastante ligações fracas, pode ser utilizada para conseguir informação sobre o perfil ou a satisfação dos clientes, por exemplo.

2.1.3 O Twitter

O Twitter é uma ferramenta de microblog lançada no final de 2006 na qual os usuários publicam mensagens com tamanho limitado de até 140 caracteres. A importância do Twitter se dá pelo seu crescimento mundial nos últimos anos atingindo a marca de mais de 200.000.000 usuários em 2011

[4] Exemplo é o caso da estudante Luiza Rabello (a Luiza que estava no Canadá), entre outros, que tornou-se sucesso em mídias como o Twitter e o Youtube e a mesma tornou-se uma espécie de celebridade instantânea, participando de programas de TV e gravando comerciais. Sobre esse caso http://f5.folha.uol.com.br/televisao/1036593-luiza-volta-do-canada-como-celebridade-e-vai-ao-jornal-hoje.shtml.

[5] http://www.orkut.com

[6] http://www.twitter.com

com cerca de 150.000.000 atualizações diárias[7].

Os usuários nessa rede constroem ligações unilaterais com outros (que podem ou não ser retribuídas) - às quais se chama 'seguir'. Tais ligações são importantes pois uma vez construídas, aquele que "segue" passa a receber as mensagens publicadas por aquele "seguido". Ou seja, construir ligações, i.e. "seguir usuários", aumentam a quantidade de informação na sua rede social.

A uma mensagem do Twitter chamamos "*tweet*" (do inglês, significando chilrear). Um *tweet* pode conter diversos símbolos importantes no contexto da rede que conotam os sentimentos ou atitudes do usuário, suas relações com os outros usuários da ferramenta ou meta-informação associada a esse *tweet*.

Um exemplo desses símbolos, importante para o nosso estudo, é o *emoticon*. Um *emoticon* é uma representação gráfica de uma emoção através de caracteres (como um sorriso ":)" conotando felicidade, ou ":(" conotando tristeza). Outros importantes exemplos são as *hashtags* - palavras iniciadas por '#' - que provêm meta-informação sobre o *tweet*. Um *tweet* pode ser direcionado a um usuário ou mencioná-lo utilizando seu nome da rede (chamado de *Twitter name* ou *username* e iniciado pelo caractere @), o qual o usuário receberá em na página de seu perfil na ferramenta.

De acordo com [76], numa avaliação manual do conteúdo dos *tweets* - uma atualização de status do usuário - , o usuário brasileiro apropria-se da ferramenta muito mais para transmitir informação que engajar em conversas - cerca de 62% dos *tweets* possuem conteúdo informacional contra cerca de 48% conversacional, com 10% dos textos possuindo possuindo ambos os perfis, i.e. informacional e conversacional. Dos textos com perfil informacional, cerca de 25% possuem conteúdo opinativo, i.e. no qual o usuário explicita uma opinião ou sentimento. Caracterizando assim o Twitter como uma mídia de cunho mais informacional.

Dada importância e o crescimento do Twitter entre os *sites* de redes sociais e o alto índice de informação opinativa que é publicada no mesmo, acreditamos que o Twitter fornece um interessante objeto de estudo, enquanto fonte de dados para os métodos propostos nesse trabalho. Adicionalmente, os textos provenientes de mídias com limitação de caracteres, como o Twitter, possuem características únicas - como um estilo de escrita muito informal e com muitas abreviações- que foram ainda pouco exploradas, principalmente na língua portuguesa.

Do ponto de vista prático, a adoção do Twitter como objeto de estudo se deu também pois, atualmente, muitas empresas passaram a utilizar essa ferramenta como um meio oficial de interação com seus clientes e de análise de seu mercado consumidor - e do seu posicionamento estratégico. Acreditamos, assim, que as soluções propostas nesse trabalho vêm para auxiliar esse novo ator dentro do processo de inteligência da empresa - o analista de redes sociais - a minerar uma quantidade cada vez maior e de mais difícil manejo.

[7] Fonte: http://business.twitter.com/basics/what-is-twitter. Estatísticas de julho de 2011

2.1.4 Marcas, *Branding* e *WOM*

Para Aaker [1], uma marca é um nome ou símbolo que identifica os bens ou serviços de um vendedor e o distingue dos seus concorrentes, tendo papel de sinalizar a origem do produto e, portanto, sua qualidade. O autor argumenta que no *marketing* moderno a criação e, principalmente, a diferenciação da marca têm sido pontos principais. Isso se dá pois a diferenciação da marca é um importante ativo da empresa que identifica qualidade ou atribui valor simbólico ao produto ou serviço, propiciando assim uma diferenciação vertical à empresa, i.e. uma vantagem competitiva sustentável.

Assim como os ativos tangíveis da empresa, entretanto, uma marca deve também ser gerenciada e passar por manutenção, para evitar que as associações feitas à ela não se deteriorem [1]. Um ponto importante para medir a lealdade do mercado consumidor à marca é a satisfação e insatisfação quanto ao produto/serviço e do sentimento associado à marca.

Nesse contexto, o conceito de *word-of-mouth* (WOM) - que pode ser entendido como "marketing boca-a-boca" em Português, definido como "comunicação interpessoal entre consumidores a respeito de uma organização ou produto" [79] ou ainda "mensagem sobre uma organização ou produtos/serviços dela" transmitido de pessoa a pessoa [13] - encaixa-se como o tipo de informação que procuramos para analisar a opinião do mercado quanto a uma marca.

Richins *et al.* [79] consideram que o WON negativo pode ter sérias consequências para uma empresa e que esse tipo de propaganda pode ter mais efeito que uma campanha de *marketing* tradicional. Charlett *et al.* [13] argumentam que o WOM é virtualmente invisível para as organizações, pois as taxas de reclamação recebidas pela empresa não representam a quantidade de insatisfações dos consumidores, uma vez que menos de um terço do consumidor comum reclama diretamente à empresa.

Jansen et al. [37] discutem a importância de WOM na internet e analisam a utilização do Twitter como fonte de WOM eletrônico. Essas fontes são importantes pois podem exploradas pela organização para descobrir a opinião do mercado quanto a seu produto ou serviço e o sentimento associado à sua marca. Uma empresa que detecte uma insatisfação pode assim agir para aumentar a confiança na marca e a satisfação associada a ela.

As mídias sociais são ambientes naturais, nesse contexto, para uma organização buscar a satisfação de seus consumidores. De fato, é isso que Pompéia [67] defende ao declarar que a empresa deve escutar os consumidores dentro das mídias sociais.

Note, porém que essa é uma tarefa difícil se considerarmos a quantidade de meios a se procurar e o volume de dados a se analisar diariamente. Acreditamos que a aplicação da Análise de Texto pode auxiliar esse processo nas empresas, diminuindo a quantidade de análise necessária e o tempo despendido nela. De fato, Jansen *et al.* [37] exploram a utilização da técnica de Análise de Sentimentos no Twitter para detecção de WOM.

2.2 Análise de Sentimentos ou Mineração de Opinião

A Análise de Sentimentos ou Mineração de Opinião, que está inserida no tópico de análise de subjetividade, corresponde ao problema de identificar (ou extrair) emoções, opiniões ou pontos de vista em textos e vem recebendo bastante atenção nos últimos anos devido à potencial aplicabilidade de tais métodos, de acordo com Wilson *et al.* [101], Akkaya *et al.* [3], Liu [50] entre outros.

Alguns teóricos identificam a atividade de análise de sentimentos com o que pode ser chamado de classificação de sentimentos, ou classificação de polaridade dos sentimentos (positivo, negativo, neutro, etc.), como exposto na definição de Liu [50]. Entretanto, os estudos de análise de sentimentos não se resumem à classificação de sentimentos.

Exemplos de trabalhos que transpõem a classificação são os de detecção de subjetividade, ou seja se uma determinada parte de um texto ou discurso possui conteúdo opinativo, como o trabalho de Pang e Lee [63]; identificação de intensidade de emoções e *flames*[8], como Wilson *et al.* [101]; identificação de pontos de vista, como em Lin *et al.* [49], ou Wiebe [97]; entre diversos outros: sumarização de opiniões [36]; viés em sumarização [83]; ou aplicação a sistemas de respostas a perguntas [92].

O problema de análise de sentimentos diferencia-se da classificação de tópicos e mostra-se difícil de tratar com as técnicas clássicas para essa área, utilizando técnicas como *bag-of-words* somente, como exposto em [65]. Diversos sistemas propostos utilizam somente informação lexical, como [33, 93], ou podem utilizar informações como a morfológica, como [65], ou ainda sintática, e.g. [101, 103].

As soluções de análise de sentimentos foram aplicadas, na literatura, a diversos problemas: desde mineração de opiniões sobre um determinado produto em blogs e fóruns [31, 33]; análise automática de resenhas de filmes em sítios como IMbB[9] [65]; de resenhas de produtos em sítios como o Amazon[10] [68]; até auxílio a Sistemas de respostas a perguntas [92] e extração de informação [81].

2.2.1 O conceito de sentimento da Análise de Sentimentos

A noção de sentimento ou opinião trabalhada pela Análise de Sentimentos é diversa e tem ligação com a terminologia utilizada pelos pesquisadores para delimitar a área.

Para Wiebe e seu grupo [98, 102], a análise de sentimentos lida com a detecção dos estados privados, ou estados internos que não podem ser observados por outros. O conceito de sentimento para esses autores está então associado à noção de estado privado, ou de estado interno do autor, de acordo com a definição de Banfield [5].

Roman [83] apresenta diversas definições para sentimentos e emoções, com origem nos campos da psicologia e neurobiologia para sentimentos. Nelas pode-se distinguir emoções de sentimentos, sendo as segundas as justaposições das alterações no estado corpóreo justaposto à imagem mental

[8] Mensagens "inflamadas"- com conteúdo ofensivo.
[9] *Internet Movie Database* - disponível em http://imdb.com
[10] http://www.amazon.com

do que ocasionou tal mudança. O sentimento consiste no sentir uma emoção. Nessa linha, não faz sentido falar de análise de sentimentos em texto, mas antes análise de emoções.

Liu [50], por outro lado, define uma opinião consistindo de uma atitude - expressa através de um determinado termo polarizado - associada a um aspecto - ou atributo - de uma entidade por um indivíduo. Uma opinião é então, por natureza, relacional, pessoal e explícita. O autor lança mão do conceito de emoção como "sentimentos e pensamentos subjetivos" [50, p.5. Tradução nossa] [11] para embasar sua noção relacional de opinião. Dentre as opiniões, para Liu, distinguem-se dois tipos: as diretas e as comparativas. As diretas associam diretamente uma emoção ou atitude a um atributo de uma entidade; as comparativas, por outro lado, "expressam uma relação de similaridades e diferenças entre dois ou mais objetos e as preferências do autor" [50, p. 5].

Outra importante distinção que este autor faz é de opinião explícita e implícita. A primeira envolve a expressão de uma opinião através de uma sentença subjetiva, i.e. através da expressão de emoções. A segunda, por sua vez, envolve a emoção implicada por uma sentença objetiva. A opinião implícita envolve, então, um tratamento também extra-linguístico e contextual - ao necessitar identificar que um determinado evento ou fato é indesejável.

Nesse trabalho, seguimos a trilha de [40] e identificamos sentimentos e opiniões com qualquer declaração de avaliação sobre um objeto ou entidade. Note então que nossa definição de opinião - ou sentimento - diz respeito a contextos avaliativos explícitos.

2.2.2 Construção de Léxicos de Opinião

Um importante recurso para a Mineração de Opinião são os Léxicos de Opinião, que consistem de listas de palavras ou expressões anotadas com informação de polaridade das mesmas. Tais recursos são utilizados em diversos métodos da literatura e para a implementação de um método de Análise de Sentimentos em uma língua, são de grande ajuda.

Trabalhos sobre a determinação da orientação semântica de palavras ou termos usualmente recaem sobre três métodos: os baseados em grandes corpora, os baseados em recursos lexicais - como dicionários ou *thesauri* - e os baseados em multi-língua ou tradução.

Os primeiros utilizam relações encontradas ente palavras e expressões presentes nos corpora para determinar sua polaridade. Trabalhos como [34, 80, 96] caem nessa categoria. Sua vantagem reside na possibilidade de identificar expressões multi-palavras opinativas como "pé no saco", expressões com polaridade adquirida por uso social e não necessariamente listados em recursos lexicográficos como "fantástico". Os resultados, entretanto, refletem diretamente a natureza do corpus utilizado e, portanto, diferentes sentidos para um palavra ou expressão pode não ser capturados e a abrangência desse léxico pode ser diminuta. Tais métodos também requerem uma grande quantidade de processamento.

A segunda abordagem explora as relações semânticas presentes em recursos lexicais como *thesauri* e dicionários. Representante distinto desses métodos é o trabalho de Kamps et al. [38] que usa a

[11] "Emotions are our subjective feelings and thoughts".

WordNet para identificar a polaridade de adjetivos ou o de Esuli et al. [23] que utiliza dicionários, entre outros. A vantagem de tais métodos está na possibilidade de explorar relações semânticas, manualmente codificadas e avaliadas, existentes entre as palavras e pela grande cobertura do léxico da língua em questão. Entretanto, tais métodos só capturam o significado lexicográfico das palavras e não são capazes de descobrir gírias ou expressões multi-palavras.

Finalmente, os métodos multi-língua e baseados em tradução exploram recursos lexicais já disponíveis para outras línguas, como o inglês, para a criação de um tal recurso para a língua alvo. Sua vantagem reside no fato de algumas línguas não possuírem recursos linguísticos construídos que permitam a utilização de outros métodos, mas devem lidar com a complexidade envolvida na tradução entre duas línguas diferentes. Exemplo desse método é o trabalho de Mihalcea *et al.* [54] que utiliza um dicionário bilíngue e um léxico de opiniões do inglês para gerar um léxico de opiniões para o romeno.

2.2.3 Análise de sentimentos granular

Métodos focados em uma análise de sentimentos mais granular - seja a nível de sentença ou de sintagma ou expressão - sugiram na literatura pela sua importância para aplicação em outros métodos como sumarização de opiniões [31, 36], perguntas e respostas [91] ou visando melhorar os resultados de métodos documentais - [63] por exemplo, faz classificação de subjetividade em sentenças para sumarizar o texto e, posteriormente, realizar a classificação de polaridade, com resultados equiparáveis aos métodos usando o texto completo.

Diversos métodos foram utilizados para resolver o problema de identificar a polaridade associada a uma sentença ou expressão. Trabalhos como [100, 103] utilizam aprendizagem de máquina para determinar a polaridade associada a um sintagma (*phrase* em inglês). Outros trabalhos utilizam regras ou heurísticas para determinar a polaridade do sentimento associado a uma sentença ou suas partes. Trabalhos como [31] aplicam diretamente a soma algébrica das polaridades das palavras constituintes da sentença - numa forma similar à feita em análises a nível documental, e.g. [96]. Trabalhos como [11, 19, 20, 41] utilizam-se de heurísticas ao explorar regras sintáticas, semânticas e discursivas manualmente codificadas. Outros, como [14, 55] utilizam uma abordagem baseada na semântica composicional, traduzindo a estrutura sintática das sentenças em formas semânticas e utilizando essas últimas para determinar a polaridade através de relações como composição, reforço, contradição, etc.

Choi e Cardie [14] mostraram que um tratamento mais apurado da negação tem grande impacto em um método granular de análise de sentimentos, em contraste à modesta melhora na performance vista em um nível documental [65]. Em [99], Wiegand *et al.* discutem que isso se dá, pois, num nível documental, a informação geralmente é redundante e mesmo em métodos simples como *bag-of-words* a polaridade geral do texto pode ser alcançada por outros itens.

Nesse contexto, é importante mencionar a dificuldade em se tratar a negação no Português Brasileiro. Nossa língua, diferente de grande parte das demais, possui três formas de negação verbal [86], a saber: uma canônica, pré-verbal, em que a partícula negativa vem antes do verbo

como em "eu **não quero** dormir!" e duas não canônicas, uma pós-verbal, em que a partícula de negação localiza-se após o verbo como em "**quero** dormir **não**!", e uma dupla, na qual o verbo é cercado por duas partículas de negação, uma antes e uma após ele, como em "**não quero** dormir **não**!". Todas as três formas possuem o mesmo significado, de acordo com Schwenter [86], e não ocorre o cancelamento da negação no caso da negação dupla. As regras usualmente associadas à negação devem então ser avaliadas para o caso do Português Brasileiro para se aplicar propriamente esses métodos à nossa língua.

2.2.4 Mineração focada em atributos e entidades

Alguns trabalhos mais recentes tem como foco o problema de determinar a opinião associada a uma certa entidade ou atributo de uma entidade. Exemplos de trabalhos com esse enfoque são [20, 31, 42, 103]. Eles procuram identificar o referente de uma determinada opinião, ou seja, a entidade à qual a opinião se refere, estabelecendo quais os fatores importantes para tal avaliação, ou seja quais atributos da mesma estão sendo julgados.

Para esclarecer, num contexto de resenhas de filmes (ou produtos) é comum aparecer uma avaliação como a seguinte:

> *"Um dos filmes mais importantes da história do cinema.[...] Cesare é um personagem que consegue passar uma forte impressão sem sequer abrir a boca durante todo o filme. Seu olhar é congelante, seus movimentos frios. [...] Embora tenha personagens notáveis, o maior marco do filme foi ter inaugurado no cinema o movimento conhecido como "Expressionismo Alemão" [...] Outra característica marcante do filme é sua excitante trilha sonora."*

Numa mesma resenha aparecem avaliações sobre o filme, sobre a trilha sonora e sobre uma personagem em particular. Tais avaliações podem variar em intensidade e polaridade, o que levaria a erros se tratadas de forma homogênea. Em casos de avaliação de produtos, um usuário pode valorizar mais uma característica do produto que outro. Há assim a necessidade de se identificar as entidades (ou características) às quais uma sentença subjetiva se refere e associar a avaliação àquela.

O método utilizado para relacionar uma entidade mencionada no texto com a opinião associada a ela varia bastante na literatura. Trabalhos como [31, 36, 40] consideram o escopo de uma expressão avaliativa como uma janela de tamanho pré-definido - quantidade de palavras, uma sentença completa, etc. - e sua polaridade é associada a toda entidade presente nessa janela. Outros, como [11,19,20,41,103], utilizam regras pré-definidas ou recursos como a FrameNet explorando a estrutura sintática ou discursiva do texto para determinar o referente de uma dada expressão opinativa. Outros ainda, como [42,43,68] utilizam um método baseado em aprendizagem de máquina para realizar a associação - estratégias fortemente relacionadas com resolução anafórica - i.e. identificação de refentes de anáforas.

2.2.5 Mineração de Twitter

Devido às apropriações realizadas da ferramenta no Brasil (c.f. 2.1.3) e no mundo [37], o Twitter vem sendo bastante explorado tanto no campo da antropologia da rede e netnografia (e.g. [29, 76]) quanto na análise de mercado por empresas.

Jansen *et al.* [37], por exemplo, discutem o Twitter como um importante recurso para empresas reconhecerem a resposta do mercado consumidor. O autor sugere e avalia aplicação de Mineração de Opinião para monitoramento de marcas no Twitter, entretanto a ferramenta relatada pelo autor não foi encontrada para uso ou descrita na literatura.

No campo da Análise de Sentimentos, entretanto, essa é uma fonte de dados ainda pouco explorada - possivelmente devido a sua recente popularidade. Muito recentemente trabalhos focados em Mineração de Opinião no Twitter vêm surgindo na literatura.

Trabalhos tratando de Análise de Sentimentos em dados do Twitter usualmente recaem na aplicação de métodos de aprendizagem de máquina, usando uma abordagem *bag-of-words* ou explorando n-gramas, como [8, 27, 28, 62]. Outros atributos - baseados em anotação POS (do inglês *Part-of-Seech*) e padrões textuais ou padrões conversacionais - são explorados em trabalhos como [6, 16, 44].

Para treino usualmente utiliza-se uma grande quantidade de *tweets* - em alguns casos centenas de milhares de exemplos. Tais métodos exploram a possibilidade de se construir automaticamente corpora de *tweets*. Esse é um importante tópico ao se tratar do Twitter, pois, diferente de outros tipos de texto, não existe ainda uma base anotada que possa ser utilizada como *benchmark* para essa tarefa. Muitos dos trabalhos sobre análise de sentimentos em *tweets* utilizam-se de marcações feitas pelos usuários - próprias dessa mídia - como *emoticons* [28, 62] e *hashtags* [16, 44]. Outra possível abordagem é a utilização de serviços de *crowdsourcing*, como o Amazon Mechanical Turk [12], para anotar o sentimento associado ao texto [17].

2.3 Reconhecimento de Entidades Nomeadas

O Reconhecimento de Entidades Nomeadas (REN ou NER, do inglês *Named Entity Recognition*) é um problema proposto na ocasião da sexta *Message Understanding Conference* (MUC-6) [32] e corresponde a identificar e classificar entidades mencionadas em textos através de designadores rígidos como nomes próprios, expressões temporais e espécies biológicas [58]. Santos e Cardoso [12] explicitam que esta tarefa é um primeiro passo na análise semântica de um texto.

Mazur e Dale [15] afirmam que o problema de reconhecimento de entidades nomeadas já está bem posicionado, tendo recebido bastante atenção da comunidade acadêmica, com técnicas do estado-da-arte obtendo alta performance. De fato, Nadeau e Sekine [59] analisando as tendências na área em 15 anos de desenvolvimento (1991 - 2006) toma um conjunto de 100 artigos escritos em língua inglesa publicados nas maiores conferências da área, um número expressivo, considerando que somente a partir de 1996 com a ocorrência da MUC-6 as publicações tratando de entidades nomeadas cresceram.

[12] https://www.mturk.com/mturk/welcome

É patente que os sistemas de REN passaram, nos últimos anos, a utilizar técnicas de aprendizagem de máquina tornando esse método o principal na área, em contraste com os primeiros sistemas propostos para tal problema que utilizavam regras manualmente codificadas e heurísticas [15,58,59].

O reconhecimento de entidades nomeadas, assim como outros problemas de reconhecimento de pedaços do discurso, como a tarefa de identificação de sintagmas nominais, recentemente vem sendo codificado como uma tarefa de descoberta sequencial de etiquetas. Existem diferentes formas de se codificar essa tarefa como uma etiquetamento sequencial, as mais populares são a codificação BIO e a BILOU [70].

Na etiquetação BIO (em inglês *Begin, Inside, Outside* ou Início, Dentro, Fora) marca-se uma determinada palavra em um texto como estando no início (B) de uma entidade nomeada, dentro (I) de uma, ou não pertencendo a uma entidade nomeada (O). O esquema de anotação BILOU (do inglês *Begin, Inside, Last, Outside, Unit*, ou Início, Dentro, Fim, Fora, Unitário) acrescenta ainda as etiquetas para a última palavra de uma entidade nomeada (L) e para entidades constituídas por uma só palavra (U). Pode-se ainda acrescer a esse esquema uma anotação quanto ao tipo de entidade criando as etiquetas B-PERSON, I-PERSON, B-LOC, I-LOC, etc.

Ratinov e Roth [70] em seus experimentos, destacam que o esquema BILOU foi mais eficiente - proporcionando melhores resultados - uma vez que representa uma codificação mais granular das informações. Técnicas computacionais usualmente associadas à utilização dessa codificação são classificadores baseados em probabilidade como o classificador CRF (*Conditional Random Fields*) [45]. Os mesmos autores ainda destacam a importância de se utilizar informação não-local, i.e. fora do simples escopo do texto ou da sentença, para garantir melhores resultados.

2.3.1 Classificação de Entidades Nomeadas

Além do reconhecimento de entidades nomeadas, uma importante tarefa associada é a categorização das mesmas, i.e. a classificação desses nomes em categorias semânticas pré-especificadas, tarefa essa já especificada desde a realização da MUC-6. No âmbito do Português Brasileiro (PT-Br) diversos sistemas foram propostos e avaliados para a solução de tais problemas, pela ocasião do HAREM - uma avaliação conjunta para o reconhecimento de entidades mencionadas. Nessa avaliação, o sistema PALAVRA-NER, proposto por Bick [7], tirou a melhor colocação.

Bick [7] e Santos [84] discutem que existem duas formas de se realizar a categorização de entidades nomeadas (mencionadas, na terminologia de Santos) baseados nas estratégias para lidar com a vagueza e a metonímia (figura de linguagem em que uma expressão é utilizada para referir outro referente relacionado), como na expressão a seguir:

- Vai querer o que, *Paraíba*?

Paraíba, nesse contexto, não se refere ao estado brasileiro da Paraíba, mas sim a uma pessoa (provavelmente proveniente desse estado). As duas formas que ambos discutem são: a lexemática, ou "forma antes da função", e funcional, ou "função antes da forma". A primeira consiste em tomar

como o padrão o sentido canônico, lexicográfico, ou seja o sentido presente nos dicionários. A segunda forma, leva em consideração o sentido no contexto.

Para o problema de identificação e classificação de ENs, mais de uma conferência visando a avaliação de sistemas foi realizada. Dentre elas, é importante, no nosso contexto, citar três: a MUC, a ACE e o HAREM.

2.3.2 MUC

A *Message Understanding Conference* (MUC) foi a primeira conferência a propor a tarefa de reconhecimento de entidades nomeadas. Nas definições da MUC-6 [32], o REN é classificado como "uma tarefa de uso prático, largamente independente de domínio e que poderia ser realizada automaticamente em um futuro próximo" [32, p.416] e corresponde à tarefa de "identificar os nomes de todas as pessoas, organizações e lugares geográficos nos textos" [32, p.416] (tradução nossa).

As categorias semânticas dessa tarefa, descritas em [94], eram a ENAMEX (*entity name expression*, em português expressão de nome de entidade), a NUMEX (*numeric expression*, em português expressão numérica), e a TIMEX (*time expression*, em português expressão temporal) que por sua vez eram subdivididas em *Person* (Pessoa), *Organization* (Organização) e *LOCATION* (Lugar), para ENAMEX, e *Money* (Dinheiro) e *PERCENT* (Porcentagem). Nota-se que, devido ao escopo de reconhecimento limitado, aquelas entidades nomeadas que não pudessem ser classificadas entre essas categorias não deviam ser identificadas, i.e. a MUC preocupava-se somente com um subconjunto das entidades presentes nos textos.

2.3.3 ACE

A ACE - *Automatic Content Extraction* - é um conjunto de avaliações com objetivo de "desenvolver tecnologias de entendimento da linguagem humana que provejam detecção automática e reconhecimento de informações-chave sobre entidades do mundo real, relações e eventos em textos de linguagem fonte e convertê-los para uma forma estruturada" [60, tradução nossa]. Em termos gerais, o ACE tem como objetivos atacar os mesmos problemas que a MUC [21].

A principal diferença entre as duas avaliações é que as tarefas do ACE são mais ambiciosas que da avaliação anterior. Por instância, o ACE define uma tarefa de reconhecimento de entidades, mas diferentemente da conferência que a precedeu, o ACE não se limita a entidades nomeadas, mas trata também de descrições de entidades e de pronomes.

Além disso, definiu-se uma nova tarefa: detecção de relações entre entidades. De fato, essa atividade já existia na MUC, entretanto de forma mais limitada, com a detecção de co-referências. Exemplos de relações entre entidades são as de co-referência, i.e. duas entidades denotam o mesmo objeto (e.g. Carmem Miranda e Pequena Notável), as de PARTE-TODO, onde uma entidade é parte de outra (e.g. Rio Grande do Sul é parte do Brasil), etc.

As categorias de entidades no ACE são: FAC (*Facility*, em português infraestrutura), GPE (*Geo-Political Entity*, em português entidade geo-política), LOC (*Location*, em português localidade), PER

(*Person*, em português pessoa), ORG (*Organization*, em português organização). Semelhantemente à MUC, no ACE as entidades que não pudessem ser classificadas em uma das categorias previamente listadas não deveriam ser reconhecidas.

2.3.4 HAREM

O HAREM - uma avaliação conjunta para o reconhecimento de entidades mencionadas - é, como expresso no próprio nome uma avaliação conjunta para o problema das entidades nomeadas em língua portuguesa. Sobre a terminologia, Santos e Cardoso [85] explicitam que "entidades mencionadas" é uma tradução do termo *named entities* em inglês, mesmo termo da qual se deriva "entidades nomeadas" utilizadas nesse trabalho. A diferença do HAREM para as outras avaliações conjuntas vistas, além de ser específica para a língua portuguesa, é o interesse em todos os nomes próprios, não somente naqueles com classes pré-definidas, e por sua abordagem funcional do problema, como já discutido anteriormente.

O HAREM possui 9 categorias, subdivididas ainda em 41 subcategorias [7, 85], de modo que serão listadas aqui somente suas categorias iniciais: PESSOA, ORGANIZAÇÃO, TEMPO, ACONTECIMENTO, COISA, LOCAL, OBRA, ABSTRAÇÃO e VALOR. Todas as entidades nomeadas detectadas que não se enquadrem em qualquer das categorias deve ser classificada como VARIADO. Na segunda edição do evento, chamada Segundo HAREM, acresceu-se à avaliação a tarefa de reconhecimento de relações semânticas entre entidades.

2.3.5 REN e Mineração de Opinião

Alguns trabalhos em Mineração de Opinião focada em entidades e atributos utilizam entidades nomeadas para identificar a referência de uma dada opinião. Diferentes métodos são explorados nesses trabalhos, assemelhando-se pouco aos métodos usuais para REN. Ding *et al.* [20] utilizam regras de associação sequenciais (através de *sequential pattern mining*) para identificação de entidades. Li *et al.* [48] exploram Conjuntos Bayesianos para decidir se uma dada palavra marcada como nome próprio ou nome próprio plural é uma entidade nomeada.

2.3.6 REN no Twitter

Recentemente, alguns trabalhos tomam como enfoque a identificação de entidades em textos do Twitter. Esses textos, como já discutido, impõem certa dificuldade em seu processamento e os métodos já aplicados em outros tipos de texto podem não funcionar muito bem para eles. Locke [52] mostra que um anotador do estado-da-arte, o Stanford NER, tem sua performance bastante reduzida quando aplicado aos textos do Twitter.

Uma dificuldade imposta por tais mensagens aos identificadores de entidades nomeadas reside no tamanho dos textos - limitados por 140 caracteres, o que, em geral, resume-se a pouco mais que uma sentença. Textos tão curtos oferecem pouca informação contextual que pode ser utilizada para a identificação das entidades. Além disso, características muito importantes para identificação de

nomes próprios em texto - como a capitalização - não são consistentemente seguidas em meios como o Twitter. É muito comum a subcapitalização - na qual palavra alguma é capitalizada no texto - e a supercapitalização - em que várias ou todas as palavras são capitalizadas, em geral procurando denotar intensidade. Assim, geralmente, os métodos devem ser adaptados para esse meio.

Tais trabalhos, em geral, utilizam técnicas do estado-da-arte associadas a estratégias de otimização, como heurísticas de pre-processamento [82] ou métodos semi-supervisionados [51] que exploram a grande quantidade de dados não-anotados disponíveis.

3. Trabalhos Relacionados

Dado o contexto geral das áreas de estudo que nos interessam nesse trabalho, passamos a analisar aqueles trabalhos de alguma forma relacionados com o que descrevemos nessa dissertação. Dessa forma, nesse capítulo apresentamos os trabalhos relacionados ao nosso estudo, de acordo com cada método utilizado no processo, discutindo decisões de construção que tomamos para o nosso método. Tal análise se dará em duas frontes: Análise de Sentimentos e Reconhecimento de Entidades Nomeadas. Assim o fazemos pois cada um desses problemas já foi extensamente estudado na literatura e a análise de cada problema separadamente nos permite entender melhor o problema total ao qual nos atemos.

3.1 Análise de Sentimentos

São vastos os trabalhos na área de Análise de Sentimentos que povoam a literatura mais recente em Processamento de Linguagens Naturais. Diversos eventos têm colocado esse problema como uma trilha de pesquisa em sua programação, mostrando a relevância dele para a área.

Dentro da miríade de métodos e enfoques existentes, nosso corte reside nos trabalhos que enfocam no processamento de dados do Twitter ou que identificam a entidade à qual é dirigida uma opinião. Portanto, dividimos essa seção em duas partes: uma sobre os métodos de AS focados em entidades, para apresentar os métodos de associação entidade-opinião investigados na literatura e outra tratando da Análise de Sentimentos em *tweets*.

3.1.1 Análise de Sentimentos focada em entidades

Como já explanado, as técnicas aplicadas para relacionar uma entidade mencionada em um determinado texto com a opinião associada a ela variam bastante na literatura.

Trabalhos como [31, 36, 40] consideram o escopo de uma expressão avaliativa como uma janela de tamanho pré-definido - quantidade de palavras, uma sentença completa, etc. - e sua polaridade é associada a toda entidade presente nessa janela. Tal estratégia é mais simplista, e admite implicitamente que qualquer entidade que apareça num contexto avaliativo está sendo avaliada. Isso não é necessariamente verdade, uma vez que tal entidade pode estar somente fracamente relacionada ao objeto de avaliação como no caso da sentença abaixo:

> "fui com meus irmãos e etc sair pra comer né, ai meu irmão levou o **ipad** dele e eu fiquei jogando **angry birds** enquanto eles conversavam, AMEI"(sic)

Nesta sentença, a expressão "AMEI" expressa uma opinião positiva sobre a entidade "angry birds" - um jogo disponível para diversos dispositivos móveis como o Ipad, mencionado na sentença. Apesar da entidade "ipad" ser relacionada com a entidade avaliada, a opinião expressa não relaciona-se com esta primeira, logo, como pode ser visto, tal estratégia pode gerar vários erros. De fato,

essa estratégia é a utilizada pela maioria das ferramentas comerciais avaliadas por nós, levando-as a uma grande imprecisão.

Outros trabalhos, como [19, 20, 41, 89, 103], utilizam regras pré-definidas ou recursos como a FrameNet - explorando a estrutura sintática ou discursiva do texto - para determinar o referente de uma dada expressão opinativa. Apesar de interessante, estratégias que utilizam informação linguística mais profunda - tal como a sintaxe ou a estrutura discursiva do texto - são dificilmente adaptáveis ao Twitter, sem que haja uma grande simplificação, uma vez que anotar tais informações seria bastante difícil, postas as características do texto. Os trabalhos de Ding *et al.* [19,20] são os que parecem melhor se adaptar ao nosso caso, por também trabalharem com textos com bastante ruído - os autores mineram fóruns de discussão sobre produtos - e por basearem-se em regras manualmente construídas explorando informação morfológica e pseudo-sintática.

Outros ainda, como [18, 42, 43, 68] utilizam um método baseado em aprendizagem de máquina para realizar a associação. Tais técnicas assemelham-se às utilizadas no problema de resolução de correferências nominais - de fato, [42] utiliza um sistema de resolução de correferências para identificar a referência de uma dada opinião e [18] explora atributos comumente usados em sistemas de resolução de anáforas para reconhecer a referência de uma opinião.

3.1.2 Análise de Sentimentos em textos do Twitter

Alguns trabalhos recentes dedicam-se ao problema de Análise de Sentimentos com um enfoque em um gênero textual específico: as mensagens do Twitter. Entre eles, podemos listar os abaixo discutidos.

Go *et al.* [28] usam unigramas e bigramas como atributos, modelando a negação como um atributo binário, indicando a presença/ausência de um negador na sentença. Os autores comparam dois classificadores (Naïve Bayes e Máxima Entropia) para predizer a polaridade das mensagens.

Pak e Paroubek [62] treinam um classificador multinomial Naïve Bayes usando n-gramas e anotação POS como atributos - usando o TreeTagger para anotar os textos. A negação é tratada com uma inversão de polaridade dentro de uma janela pre-definida de três palavras - da imediatamente anterior à expressão de negação, à imediatamente posterior.

Bifet e Frank [8] usam unigramas como atributos de classificação e comparam três diferentes classificadores. Os autores argumentam que como os dados do Twitter possuem uma característica de *stream* - dispersos ao longo do tempo - a métrica de acurácia é inadequada para avaliação dos resultados de classificadores sobre esses dados. Propõem então a utilização da estatística *Kappa* sobre uma janela deslizante de textos em períodos de tempo, o que permitiria uma caracterização melhor do desempenho dos classificadores em relação ao sentimento dos textos dentro daquele *spam* temporal.

Davidov *et al.* [16] exploram o uso de *hashtags* e *emoticons* como rótulos de sentimentos - i.e. como os sentimentos que o sistema utilizará para classificar os *tweets* - e treinam um classificador KNN (do inglês *K Nearest Neighbors*, K vizinhos mais próximos) usando n-gramas e atributos baseados em padrões. Kouloumpis et al. [44] usam uma abordagem similar, entretanto realizam a

classificação para as polaridades do sentimento expresso no texto, i.e. classificam-no como positivo, neutro ou negativo.

Barbosa e Feng [6] usam anotação POS, presença de maiúsculas nas palavras, *emoticons*, pontuação e atributos específicos do Twitter como presença de links no *tweet*, *Twitter names* e marcadores RT - marcadores que indicam que a mensagem está sendo retransmitida, i.e. que a mesma é de autoria de outrem. A classificação é realizada em dois passos: no primeiro eles classificam os textos como subjetivos ou objetivos, i.e. conotando um sentimento ou não; no segundo passo os *tweets* classificados como subjetivos são classificados de acordo com a polaridade do sentimento transmitido.Os autores utilizam um classificador SVM (do inglês *Support Vector Machines*, máquinas de vetor de suporte) para ambas as etapas.

Esses trabalhos tratam de classificar um determinado texto do Twitter quanto ao sentimento transmitido. Nosso enfoque, entretanto é classificar o sentimento associado a uma determinada entidade. Relacionado a esse nosso objetivo, são os trabalhos de Jansen *et al.* [37] e Silva *et al.* [89].

Jansen *et al.* [37] usam uma aplicação comercial - atualmente indisponível - para realizar a análise de sentimentos associados a uma marca no Twitter, possibilitando assim a detecção de WOM no Twitter.

O trabalho de Silva *et al.* [89] descreve a construção de uma ferramenta que utiliza análise de sentimentos para descobrir as opiniões associadas aos candidatos da eleição presidencial portuguesa expressas no Twitter. Eles utilizam uma estratégia baseada em léxico combinado com regras lexico-sintáticas para identificar e compor os sentimentos e atribuir a referência a uma determinada opinião.

Para o Português ainda, Silva *et al.* [87] utilizam o conceito de viés do autor quanto a um tópico e, assim, classificar uma opinião sobre o mesmo em *tweets*. Os autores trabalham no domínio da política e de esportes. Calais Guerra *et al* [10], por sua vez, utilizam regras de associação - numa estratégia similar à de [8] - para classificar *tweets* quanto a polaridade. Os autores aplicam sua técnica para detecção de casos de dengue no Brasil, mas não avaliam a performance do seu método.

Acreditamos que - apesar de tratarem problemas, de certa forma, similares ao nosso - tais estratégias não são adequadas pois não se aplicam ao caso que estudamos - viés faz sentido no domínio de política ou em contextos em que há uma grande polaridade entre duas entidades, mas não se adéqua, em nossa opinião, à analise de mercado que procura diferentes opiniões sobre um produto ou marca - e pelo baixo nível de granularidade em que se analisam as opiniões.

Neste trabalho, dadas as características do texto do Twitter e as ferramentas disponíveis, tomamos uma estratégia de mineração baseada em léxico e de associação baseada nos métodos de resolução de co-referência. Nosso trabalho é inspirado pelos trabalhos de Silva *et al.* [89] e Ding *et al.* [20], apesar destes autores utilizarem regras manualmente codificadas. Acreditamos que, dadas as poucas informações linguísticas de que dispomos, uma estratégia que explore o aprendizado automático pode extrapolar informações dos dados que seriam dificilmente codificadas em regras pré-estabelecidas.

3.1.3 Comparação entre os trabalhos

Apresentamos, na Tabela 3.1, uma comparação entre os trabalhos relacionados baseada no tipo de técnica de AS usada, o tratamento da negação, se utilizou método de aprendizado automático para determinação da polaridade, a profundidade de análise, tipo de texto (ou mídia) tratado e método de atribuição de referência.

Dentre os trabalhos que consideramos mais relacionados com o nosso, fazemos ainda uma comparação em relação à adequação às características desejáveis para tratar o problema que nós propusemos. Podemos ver na Tabela 3.2 que os outros trabalhos não se adequam completamente ao problema proposto, seja por se basear em regras específicas para determinados domínios, por não ter sido aplicado para o Português ou por tratar o fenômeno em um nível de análise que consideramos inadequado, como já discutido.

3.2 REN

Trabalhos sobre Reconhecimento de Entidades Nomeadas são diversos na literatura. Para o caso da língua portuguesa, os participantes do HAREM são de especial importância. Tais métodos, entretanto, são pouco indicados no nosso caso específico pela natureza dos textos do Twitter. O Palavras-NER - sistema que recebeu o primeiro lugar na trilha geral da primeira edição do HAREM - conseguiu identificar menos de 20% das entidades mencionadas no Twitter num conjunto de cerca de 50 textos dessa fonte anotados com ele. Por tal motivo concentramos nossa discussão nos trabalhos sobre Reconhecimento de Entidades Nomeadas focados no contexto do Twitter.

3.2.1 REN no Twitter

Locke [52] investiga a tarefa de REN em textos do Twitter aplicando um classificador treinado em textos jornalísticos - provenientes do corpus provido para treino na tarefa de REN do CoNLL-2003. O sistema - baseado em SVM - teve baixa performance nos textos do Twitter, com medida F[1] inferior a 40%. Os *tweets* utilizados para a avaliação foram manualmente anotados de acordo com os padrões do CoNLL-2003 [95].

O trabalho de Ritter *et al.* [82] utiliza uma estratégia baseada em agrupamento de palavras para realizar normalização lexical no Twitter e atributos morfossintáticos para treinar um reconhecedor de entidades nomeadas baseado no algoritmo CRF. Esses autores tratam a identificação e a classificação como tarefas separadas, aplicando um classificador *LabelledLDA* para esta última. Os autores alcançam uma marca de 67% de medida F para o reconhecimento e 66% para a classificação com a sua estratégia - sobre um conjunto de 800 *tweets*.

Finin *et al.* [25] propõem a utilização de serviços de *crowdsourcing* - como o *Amazon Mechanical Turk*, já citado anteriormente - para criar um recurso dourado com anotação de entidades nomeadas

[1] Medida ponderada entre a precisão (P) e a abrangência (R) dada por $F_\beta = (1+\beta^2) \cdot \frac{\mathrm{P} \cdot \mathrm{R}}{(\beta^2 \cdot \mathrm{P}) + \mathrm{R}}$. Tomando $\beta = 1$ temos a medida F1 com mesmo peso para Precisão e Abrangência.

Tabela 3.1 – Tabela comparativa dos trabalhos relacionados sobre Análise de Sentimentos

Trabalho	**Técnica**	**Aprendizado?**	**Tratamento de Negação**	**Profundidade de Análise**	**Mídia**	**Atribuição de referência**
Greffenstette *et al.* [31]	Léxico	Não	-	-	Resenhas	Presença no texto
Hu e Liu [36]	Léxico	Não	Janela	Sentencial	Resenhas	Proximidade
Kim e Hovy [40]	Léxico	Não	-	Sub-sentencial	Jornalísticos	Proximidade
Kim e Hovy [41]	Léxico	Não	-	Sub-sentencial	Jornalísticos	FrameNet
Ding *et al.* [19, 20]	Léxico + heurísticas	Não	Regras	Sub-sentencial	Fóruns da Internet	Heurísticas
Wu *et al.* [103]	Léxico	Não	-	Sub-sentencial	Resenhas	SVM
Silva *et al.* [89]	Léxico + Regras	Não	Regras	Sub-sentencial	Twitter	Regras
Popescu e Etzioli [68]	Regras + *Relaxation Labelling*	Sim	Regras	Sub-sentencial	Não Informado	Regras
Kobayashi *et al.* [42, 43]	Léxico + Regras	Sim	-	Sub-sentencial	Resenhas	SVM
Go *et al.* [28]	Unigramas e Bigramas	Sim	Sentença	Sentencial	Twitter	-
Pak e Paroubek [62]	N-gramas e POS	Sim	Janela	Sentencial	Twitter	-
Bifet e Frank [8]	Unigramas	Sim	Janela	Sentencial	Twitter	-
Davidov *et al.* [16]	N-gramas + Regras	Sim	-	Sentencial	Twitter	-
Barbosa e Feng [6]	POS, emoticons, pontuação, *features* do Twitter	Sim	-	Sentencial	Twitter	-
Jansen *et al.* [37]	Unigrama + Bigrama	Não Informado	Não Informado	Sentencial	Twitter	-
Silva *et al.* [89]	Léxico + Regras	Não	Regras	Sub-sentencial	Twitter	Regras
Nosso Trabalho	Léxico	Não	Janela	Sub-sentencial	Twitter	SVM

Tabela 3.2 – Trabalhos de AS diretamente relacionados ao nosso

Trabalho	Método de SA	Específico para o Twitter	Independente de domínio	Específico para a língua portuguesa
Ding *et al.* [20]	Sub-sentencial	Não	Não	Não
Jansen *et al.* [37]	Sentencial	Sim	Sim	Não
Silva *et al.* [89]	Sub-sentencial	Sim	Não	Sim
Nosso trabalho	Sub-sentencial	Sim	Sim	Sim

em dados do Twitter. Tal recurso foi, posteriormente, utilizado para treinar o Stanford NER, um identificador de entidades nomeadas baseado no algoritmo CRF de aprendizagem automática [57]. O autor avalia a qualidade da anotação e sua relação com a métrica de qualidade provida pelo serviço de anotação (WorkerRank) utilizando diversas métricas, mostrando que não necessariamente tais métricas estão relacionadas.

O trabalho de Ferragina e Scaiella [24] é, de certa forma, relacionado à tarefa proposta. Os autores propõem um sistema que anota textos do Twitter com ligações para verbetes da Wikipédia. A identificação de Entidades Nomeadas pode ser então realizada limitando-se os verbetes àqueles descrevendo entidades das categorias semânticas desejadas. A anotação é realizada a partir de métricas de grau de relação de uma âncora - um segmento de texto - com os verbetes da enciclopédia.

Liu *et al.* [51] utilizam uma estratégia semi-supervisionada e classificadores KNN e CRF para identificar as entidades em textos do Twitter. Os classificadores utilizam um conjunto diferente de atributos em seu treinamento, numa estratégia de *co-training*, levando em consideração informações locais, dadas pela vizinhança da palavra a ser anotada, e não-locais, através do uso de um *gazeteer*. O sistema atinge bons resultados, acima dos sistemas construídos para outros gêneros textuais quando aplicados nos textos provenientes do Twitter. Os autores utilizam como conjunto de anotação a estratégia BILOU juntamente com as categorias semânticas PERSON (Pessoa), ORGANIZATION (Organização) e LOCATION (Local).

3.2.2 Comparação entre os trabalhos

Apresentamos, na Tabela 3.3, uma comparação entre os trabalhos relacionados e o nosso trabalho, baseado no tipo de técnica de REN usada, os atributos analisados e tipo de texto (ou mídia) com o qual o sistema foi treinado. Note que, apesar de apresentarmos os valores de medida F alcançados pelos métodos, tal medida não pode ser tomada como base de comparação, uma vez que os experimentos foram realizados com bases diferentes e mesmo considerando conjuntos diferentes de categorias ou esquemas de segmentação de entidades. Em particular, os trabalhos de Finin *et al.* e Ferragina e Scaiella - marcados com (*) - não apresentam os resultados em termos de medida F ou não aplicam o método para o problema de REN, respectivamente.

Tabela 3.3 – Tabela comparativa dos trabalhos relacionados sobre Reconhecimento de Entidades Nomeadas para o Twitter

Trabalho	**Técnica**	**Atributos**	**Treino**	medida F
Locke [52]	SVM	Ortográficos, Lexicais, Morfológicos e Dicionários	Jornalísticos	31,5%
Ritter et al. [82]	CRF	Ortográficos, Lexicais, Morfológicos, Contextuais e Dicionários	Twitter	67%
Finin et al. [25, 57]	CRF	Lexicais e Morfológicos	Twitter	(*)
Liu et al. [51]	KNN e CRF em *co-training*	Ortográficos, Lexicais, e Dicionários	Twitter	80,2%
Ferragina e Scaiella [24]	Associação estatística	Wikipédia e Lexicais	Twitter	(*)
Nosso trabalho	KNN e CRF em série	Lexicais, Morfológicos e Dicionários	Twitter	-

4. Análise de Sentimentos focada em entidades no Twitter

Uma vez apresentados os métodos propostos na literatura para tratar os problemas relacionados ao objetivo dessa dissertação, podemos, então, apresentar nossa proposta - e defendê-la - assim como os passos necessários para a implementar.

Para guiar o desenvolvimento do trabalho e atacar cada problema separadamente, desenvolvemos um *framework* que organiza as tarefas a serem realizadas para minerar opiniões em mídias sociais dadas as restrições temporais e computacionais almejadas. Tal *framework* é descrito na Seção 4.1. Um importante recurso para a realização da Análise de Sentimentos é um léxico de opiniões, algo que não fomos capazes de encontrar para o Português Brasileiro. Uma parte do trabalho então dedicou-se à construção de um tal léxico, tópico tratado na Seção 4.2.Textos provenientes de mídias sociais apresentam um alto grau de agramaticalidade e de disgrafias de palavras. Desenvolvemos então um conjunto de heurísticas para normalizar tais textos, que serão tratadas na Seção 4.3. Por fim atacamos os problemas de Mineração de Opinião (Seção 4.4), e preliminarmente de Reconhecimento de Entidades Nomeadas em tais textos (Seção 4.5).

4.1 Proposta de *Framework*

Dada a necessidade de utilizarmos um processo com baixo tempo de resposta - um requisito muito importante para análise do contexto competitivo – e a agramaticalidade característica dos textos provenientes de mídias sociais, tais como blogs, redes sociais (como o Twitter) e fóruns online, optamos por utilizar informação linguística mais superficial, não utilizando, então, *parsers* ou analisadores mais profundos.

A partir de tais definições iniciais de objetivo do trabalho, os recursos disponíveis e as restrições impostas – como tempo de resposta e de utilização de ferramentas – propomos então um *framework* para a Mineração de Opinião centrada em entidades em mídias sociais utilizando informação linguística superficial - como informação lexical, morfológica ou pseudo-sintática - que pode ser analisado na Figura 4.1.

O primeiro passo consiste de uma normalização lexical dos textos, posto que textos provenientes de mídias sociais comumente apresentam um alto índice de variação linguística e disgrafias (c.f.

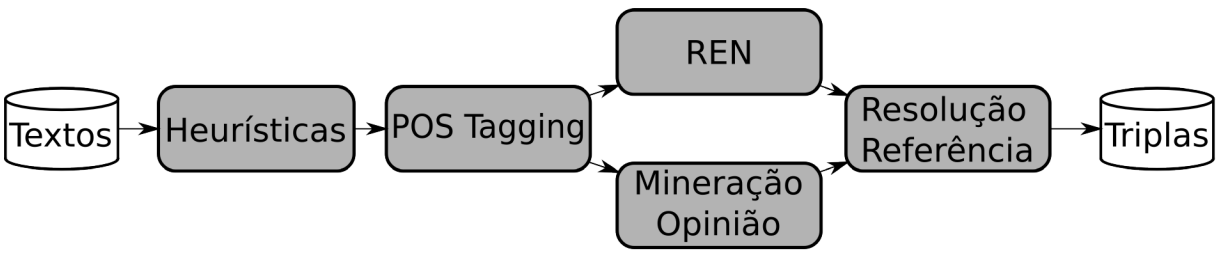

Figura 4.1 – *Framework* proposto para mineração de opiniões em mídias sociais

Seção 4.3). A segunda fase consiste de uma anotação morfológica das palavras com um etiquetador morfológico (*POS Tagger*). Para essa tarefa escolhemos o *tagger* de Brill [9], que já foi implementado e treinado para o Português [2].

Após esse passo, passamos para as fases de identificação de entidades (c.f. Seção 4.5) e Mineração de Opinião (c.f. Seção 4.4) realizadas paralelamente. Optamos por identificar essas informações separadamente pois acreditamos que um considerável esforço de pesquisa já foi realizado para resolvê-los [50, 64] podendo-se assim aproveitar os métodos já estudados na literatura e combiná-los, no lugar de desenvolver um único método que ataque ambos os problemas conjuntamente. Por fim, segue a etapa resolução de referência das opiniões que consiste de estabelecer o relacionamento entre opiniões e entidades, que implementa o processo de extração de opiniões ao relacionar com uma dada entidade a opinião associada a ela num dado texto, gerando assim triplas $< Entidade, Opinião, Texto >$.

4.2 Construção do Léxico

Um importante recurso para a realização da Análise de Sentimentos é um léxico de opiniões [50], sua importância reside em facilmente aumentar a abrangência da identificação de expressões opinativas e fornecer pistas para identificar novas, além de poder ser utilizado na identificação da polaridade das mesmas.

Não fomos capazes de encontrar para o Português Brasileiro um recurso desses que fosse independente de domínio. De fato, o único léxico de opiniões encontrado foi o SentiLex [88] criado para o Português Europeu e específico ao domínio de julgamentos sociais, ou seja, sobre pessoas.

Uma parte do trabalho então dedicou-se à construção de um tal léxico. Utilizando as técnicas baseada em corpora de Turney [96], baseada em recursos lexicais de Kamps *et al.* [38] e uma variação da técnica de Mihalcea *et al.* [54] construímos um léxico de 7077 expressões para o Português Brasileiro, anotado com a polaridade das mesmas. Como recursos para a construção desse léxico, utilizamos um corpus de resenhas de filme, criado por nós, o Thesaurus Eletrônico do Português [53] e o léxico de opiniões construído por Liu para o inglês[1] [36].

Ao utilizar métodos diferentes e explorar vários recursos, acreditamos que pode-se melhorar os resultados ao combinar as melhores características para superar as desvantagens de cada método. Assim, ao integrar, por exemplo, um método baseado em corpus com um método baseado em *thesaurus* consegue-se explorar tanto as associações semânticas manualmente codificadas do segundo recurso, com as conotações sociais e termos multi-palavras que podem ser extraídos com o primeiro método.

A construção do léxico e sua avaliação, realizadas como parte integrante da pesquisa realizada nessa dissertação, foram publicadas no STIL 2011 [90]. O léxico resultante foi, posteriormente, estendido utilizando um corpus maior gerando um léxico com mais de 10.000 termos polarizados.

Posteriormente, construímos ainda um dicionário de *emoticons* e *hashtags* que denotem opinião.

[1] Disponível em http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html.

Hashtags são usualmente utilizadas como meta-informação nos *tweets*, seja para expressar uma informação pragmática em forma textual, como ironia, ou avaliação (e.g. as *hastags* #not, #win, #fail), ou para classificá-los quanto a tópico (e.g. #google, #android, etc.).

Aquelas que expressam informações pragmática são de especial interesse e foram bastante exploradas por nós (c.f. Seção 4.4) para, por exemplo, construir o corpus. As outras entretanto, também fornecem informações importantes pois a identificação de um tópico é uma importante pista para o método quanto à polaridade do *tweet* - de acordo com a polaridade média daquele tópico num certo período de tempo relevante.

Note que essa polaridade associada ao tópico é intrinsecamente temporal e, portanto, o dicionário de *hashtags* deve ser construído dentro de limites temporais rígidos, i.e. deve ser renovado, ou reconstruído após uma certa quantidade de tempo. Por sorte, pode-se fazê-lo de forma automática. Nos utilizamos nesse trabalho das etiquetas de avaliação (#win e #fail) para identificar a polaridade dos *tweets* e uma métrica de polaridade - variando entre -1 e 1 - baseando-nos no trabalho de Turney [96]. Assim, aplica-se o método de Turney - centrando as buscas no Twitter com a TwitterAPI - utilizando as etiquetas #win e #fail para identificar os *tweets* polarizados. Da mesma forma, a polaridade dos *emoticons* pode ser automaticamente calculada.

4.3 Pré-processamento lexical e morfológico

Em estudos iniciais, dado um corpus de 6129 *tweets* extraídos durante cerca de três dias utilizando a *engine* de busca do Twitter usando a *hashtag* #fato – um meme da comunidade de usuários do Brasil – contendo 113341 palavras -excluindo-se *hashtags* e menções a outros usuários - das quais 13210, cerca de 11%, não estão presentes no léxico do português DELAF com 880.000 palavras flexionadas [56]. Desenvolvemos então um conjunto de heurísticas para normalizar tais textos quanto a grafia.

A variação de grafia mais comum identificada nos textos é a repetição de vogais - utilizada para conotar intensidade. Essa variação pode ser facilmente corrigida através de expressões regulares simples. Outro caso comum foi a variação por similaridade fonética, ou seja quando o autor troca letras ou conjunções de letras que denotam fonemas parecidos intencionalmente ou não - como trocas de letras como 's' e 'ç', ou mesmo de escritas socialmente demarcadas ou típicas de grupos e socioletos próprios da Internet, como a grafia "molier" para a palavra "mulher" ou "naum" para "não". Tal variação foi tratada utilizando-se uma variação do algoritmo Metaphone adaptado para o Português Brasileiro.

Outra fonte comum de variação envolve a utilização de abreviações de palavras ou expressões. Essas abreviações são - ou parecem ser - em muitos casos aleatórias ou pouco estruturadas, como para as grafias "vdd" para "verdade" ou "lol" para "laughing out loud" ("rindo alto" ou "rindo muito", em inglês). Um caso específico de variação percebida é a tendência de omissão de vogais para reduzir o tamanho do texto. Assim, palavras como "saudade" são grafadas como 'sdd'[2].

[2]O caso da palavra "verdade" parece ser similar, com a letra "r" sendo omitida por possuir menor saliência

A heurística proposta para esses casos é uma comparação entre palavras considerando somente suas consoantes. Essa heurística deve ser utilizada com cautela por sua alta taxa de erros, e.g. "naum" e "nome" são consideradas similares por ela. Só a aplicamos no caso em que a palavra não identificada não possui vogais.

Outras variantes não contempladas ainda incluem palavras de outros idiomas que aparecem nos textos, algo que não pretendemos tratar.

Após normalizados, os textos são morfossintaticamente anotados utilizando etiquetador morfossintático baseado em transformações de Brill [9] do NLTK[3]. Como não possuímos um conjunto de *tweets* morfossintaticamente anotados em Português, não podemos avaliar a acurácia de tal anotador, entretanto, para textos jornalísticos - do corpus MacMorpho[4] - o etiquetador possui acurácia de 91% - um índice baixo para o estado da arte. Outros etiquetadores do NLTK foram testados, entretanto nenhum atingiu um resultado satisfatório.

4.4 Análise de Sentimentos

Para a identificação de opiniões expressas em texto, utilizamos um método baseado em léxico e marcadores do Twitter - com *hashtags* e *emoticons*.

Nossos experimentos iniciais exploraram as possibilidades de aplicação de diferentes métodos considerando sua complexidade. Avaliamos desde a aplicação do léxico, numa estratégia comumente usada na literatura para a AS documental - a saber a soma algébrica das polaridades das expressões encontradas nos *tweets* - até investigando o papel da negação, numa forma mais aproximada dos métodos específicos para AS sentencial e do impacto do pré-processamento nessa tarefa. O resultado desses testes iniciais será publicado no PROPOR 2012.

Dentre as variações estudadas, estão duas diferentes modelagens para o escopo da negação, baseadas numa janela de tamanho pré-definido. O escopo da negação tem forte ligação com o contexto sintático e discursivo em que aparece [14], entretanto tais informações são indisponíveis para o nosso sistema. Testamos então dois tipos de escopo mais simples para modelagem do efeito da negação em sentenças - a saber, uma negação com escopo em toda a sentença, já que o termo de negação não necessariamente deve aparecer adjacente à locução negada, e uma negação baseada na vizinhança, já que mais usualmente o termo da negação está na adjacência da locução negada. A modelagem da negação teve pouco impacto nos nossos testes (c.f. Capítulo 5), entretanto decidimos nos ater à modelagem baseada em vizinhança por acreditarmos que modela melhor o comportamento "padrão".

O algoritmo de Análise de Sentimento funciona da seguinte forma:

```
para cada palavra no tweet
   se palavra está marcada com polaridade no léxico
```

fonética.

[3] http://www.nltk.org/.

[4] Integrante do corpus NILC São Paulo - http://www.linguateca.pt/acesso/NILCsaocarlos.html - e disponibilizado com o NLTK.

```
      adicione palavra à lista de palavras polarizadas
      se existe um termo de negação no escopo da palavra
         troque a polaridade da palavra
   se palavra é uma hashtag polarizada
      adicione palavra à lista de palavras polarizadas
   se palavra é um emoticon polarizado
      adicione palavra à lista de palavras polarizadas

retorne o conjunto de palavras polarizadas
```

Utilizamos esse conjunto para identificar os termos polarizados da sentença - para os quais a referência da opinião será posteriormente inferida.

Note que a modelagem da negação apresentada contempla a possibilidade da dupla negação, pois não multiplica a quantidade de termos negativos no contexto do termo polarizado.

4.5 Reconhecimento de Entidades

Escolhemos, dentre os métodos propostos na literatura de REN, utilizar a etiquetamento sequencial para resolver o Reconhecimento de Entidades Nomeadas, seguindo a proposta de [70]. Para implementar tal método, utilizamos a linguagem de programação Python e a biblioteca de processamento de linguagem natural NLTK - que por sua vez utiliza a implementação de CRFs da *toolkit* Mallet[5].

O classificador CRF utilizado foi treinado utilizando atributos léxicos, morfológicos e baseados em *gazeteers*, mas devido a particularidades da interface do classificador no NLTK, outras informações contextuais sugeridas em [70] não puderam ser implementadas. A principal de tais limitações foi a restrição do contexto a ser analisado à sentença sendo anotado no momento. Tal restrição impede a aquisição das informações contextuais, que servem como um histórico das decisões feitas anteriormente.

Os dados anotados utilizados durante o treino são constituídos de 1000 *tweets*, coletados usando a Twitter API buscando textos em Português usando as expressões "Google", "#Cielo" e "máquina cielo". A escolha de tais expressões se deu devido ao enfoque do trabalho - i.e. a utilização de Análise de Texto com finalidade de auxiliar o monitoramento de marca na Internet - escolhendo nomes de empresas - Cielo e Google. Tais textos foram pré-processados usando as heurísticas descritas na Seção 4.3 e morfossintaticamente anotados.

Inicialmente, considerou-se implementar o método explorado por Liu *et al.* [51], um método semi-supervisionado baseado em *co-training*. Tal escolha se deve ao fato que utilizando um método semi-supervisionado, somente uma pequena quantidade de dados anotados e uma grande quantidade de dados brutos são necessários para o treino. Como dados brutos - i.e. não anotados - são fáceis de se conseguir utilizando a API do Twitter, precisaríamos então anotar somente uma quantidade

[5] http://mallet.cs.umass.edu/

pequena de *tweets*, tornando essa estratégia bastante atraente, dado que a anotação é bastante laboriosa.

Essa estrategia mostrou-se infrutífera, entretanto, pois a interface do NLTK não nos fornecia informações essenciais para implementar o algoritmo proposto por [51] - o valor da verossimilhança (*likelihood*, em inglês) de uma determinada sequência de etiquetas dado o modelo estimado. Uma outra implementação de CRFs foi testada, entretanto a mesma comportou-se pobremente - com valores muito abaixo dos conseguidos com a implementação do Mallet, o que nos levou a questionar a correteza desta implementação.

Seguindo o trabalho de Ratinov e Roth [70], optou-se por utilizar a codificação BILOU (*Begin*, *Inside*, *Last*,*Outside*, *Unit*) que representa a informação de uma palavra pertencer a uma entidade nomeada de acordo com sua posição nesta entidade - no início do nome de uma entidade com mais de uma palavra (B), no meio do nome de uma entidade com mais de uma palavra (I), no final do nome de uma entidade com mais de uma palavra (L), não ocorre em uma entidade nomeada (O) e é o nome simples de uma entidade (U).

Ao optarmos pela implementação do classificador CRF do NLTK (que utiliza o Mallet) podemos excluir a correta implementação do algoritmo como fator de bloqueio de nossos experimentos - uma vez que o mesmo já foi amplamente utilizado e testado na comunidade. Acreditamos que - apesar de reduzir a flexibilidade na construção do reconhecedor proposto - um método supervisionado baseado em CRF, dados os esforços de preprocessamento e o treinamento em dados provenientes do Twitter - e não em textos de outro gênero - nos fornece um bom modelo para Reconhecimento de Entidades Nomeadas no Twitter para o Português.

4.6 Identificação de referência

Para a identificação de referência da opinião utilizamos um método de aprendizagem de máquina sobre um conjunto de atributos similar a [18]. Os autores utilizam atributos bastante utilizados em reconhecimento automático de correferência nominal para identificar a referência de uma opinião - como concordância de gênero e numero, atributos posicionais da opinião e do candidato, etc.

O método foi implementado na linguagem de programação Python utilizando a biblioteca SciKits[6] para aprendizagem de máquina. Utilizamos, como algoritmo de aprendizagem de máquina, a implementação do SciKits de uma máquina de vetor de suporte (SVM). Como atributos fornecidos ao algoritmo de aprendizagem utilizamos:

- Distância entre o termo polarizado (termo que expressa opinião) e a entidade nomeada candidata a referência, calculado como a diferença entre o índice da posição de início do nome da entidade na sentença e do termo polarizado;

- Valor absoluto da distância entre o termo polarizado e a entidade nomeada candidata a referência;

[6] http://scikits.appspot.com/

- Índice da posição de início do nome da entidade na sentença;
- Índice da posição de início do termo polarizado na sentença;
- Quantidade de entidades na sentença
- Tamanho da sentença;
- Razão entre o índice da posição de início do nome da entidade na sentença e o tamanho da sentença - ou fator de centralidade da entidade nomeada;
- Razão entre o índice da posição de início da expressão polarizada e o tamanho da sentença - ou fator de centralidade da expressão polarizada;
- Razão entre a distância entre o termo polarizado e o candidato e o tamanho da sentença - ou distância suavizada;
- Concordância de número entre entidade e expressão polarizada: valor booleano, calculado heuristicamente pela presença do 's' ao final do nome ou da expressão.

A escolha do SVM como método de aprendizagem se deu, em grande parte, por sua baixa sensitividade a desbalanciamento entre classes, uma vez que existem muito mais casos negativos de referência entre termos polarizados e entidades que positivos.

5. Resultados e Avaliação

Nesse capítulo discutimos o processo de avaliação dos métodos propostos no Capítulo 4, assim como os dados construídos e utilizados para tanto, além de apresentar os resultados de tal avaliação.

5.1 Construção e Anotação do corpus de teste

Como recurso para avaliar o desempenho do sistema como um todo - e das técnicas individuais escolhidas - anotamos um conjunto de 1000 textos extraídos do Twitter contendo as palavras "google", "ipad" ou "nokia", obtidos utilizando-se a API do Twitter procurando por textos em língua portuguesa contendo tais expressões. Tal conjunto de *tweets* compõe o corpus TwitterSentiment de mensagens do Twitter anotadas com sentimento para o Português. A escolha das expressões foi feita com foco nos objetivos do trabalho - uma vez que nosso interesse é aplicar Análise de Sentimentos para auxiliar um processo de Inteligência Competitiva. Escolhemos então empresas e produtos da área de tecnologia e comunicação por serem frequentemente citados na mídia escolhida.

A anotação de tais textos se deu em duas partes: a anotação de entidades e a anotação de opinião e referências. Os 1000 textos foram separados em três partes para garantir uma maior agilidade na anotação. Cada parte continha 400 textos, com 100 desses em comum entre os três conjuntos. Três anotadores anotaram uma das partes cada e o conjunto de 100 textos em comum serve para podermos calcular o grau de concordância entre tais anotações. A anotação foi realizada em um ciclo, seguindo as regras no Apêndice A.

Medimos a concordância entre os anotadores através da estatística Kappa, mostrados na Tabela 5.2. A estatística Kappa pode ser interpretada como um índice de concordância entre dois classificadores - nesse caso anotadores - de acordo com Altman [4]. O autor propõe interpretação ilustrada na Tabela 5.1 para os valores da estatística Kappa:

Tabela 5.1 – Interpretação de Altman para a estatística Kappa como medida de concordância.

Intervalo de valores	Interpretação
$< 0,2$	Concordância fraca
$0,2$ a $0,4$	Alguma concordância
$0,4$ a $0,6$	Concordância moderada
$0,6$ a $0,8$	Boa concordância
$0,8$ a 1	Concordância excelente

Como pode-se ver na Tabela 5.2, a anotação de entidades nomeadas obteve uma concordância excelente.

Para avaliar a segunda parte da anotação - que consiste da identificação de termos que conotam opinião, sua polaridade e atribuir sua referência - devido a quantidade bastante reduzida de anotações considerando-se os 100 textos em comum entre os anotadores - dadas as restrições impostas nas

Tabela 5.2 – Estatística Kappa calculada sobre a delimitação de entidades nomeadas

Anotadores	Concordância
Anotador 1 X Anotador 2	0,92
Anotador 1 X Anotador 3	0,94
Anotador 2 X Anotador 3	0,94

instruções de anotação (c.f. Apêndice A) - consideramos que a estatística Kappa não é uma medida apropriada. Analisando manualmente os resultados percebemos que:

1. Os anotadores marcaram entre 5 (Anotador 2) e 8 (Anotador 3) termos polarizados referindo-se a uma entidade nomeada;

2. Dentre estes, todos concordam com 4;

3. Em todos os casos onde houve concordância da delimitação do termo polarizado entre pelo menos dois anotadores, a referência e a polaridade foram igualmente identificadas.

Acreditamos que a anotação possui concordância aceitável para o problema proposto, apesar dos poucos casos anotados no conjunto de controle não nos permitir avalitar quantitativamente sua qualidade.

5.2 Análise de Sentimentos

A avaliação do método de Análise de Sentimentos se deu em duas partes. A primeira consiste na avaliação do método para a classificação sentencial quanto à polaridade, i.e. classificar uma dada sentença como positiva ou negativa. A segunda parte, consiste na identificação de opiniões relacionadas a entidades e, portanto, sub-sentencial, pois envolve descobrir quais partes da sentença se referem a quias entidades mencionadas.

A avaliação se deu assim para primeiramente podermos avaliar a qualidade do método de classificação de sentimentos em relação aos outros da literatura, posto que a maioria dos que trabalham com o Twitter fazem a classificação de uma sentença e não focada em entidade. A segunda avaliação nos permite, então, analisar a performance do método no nível de granularidade requerido pela nossa proposta.

5.2.1 Classificação de sentenças opinativas

Uma vez que o corpus TwitterSentiment - descrito na Seção 5.1 - não possui informação de polaridade a nível de sentença, foi necessário utilizarmos outro recurso para realizar esta etapa de avaliação.

Construímos, então, um conjunto de 1700 *tweets* em língua portuguesa divididos entre as classes Positivo e Negativo - 850 para cada classe - representando a polaridade associada ao sentimento transmitido pelo texto. Os textos foram coletados automaticamente utilizando a Twitter API,

buscando pelas *hashtags* #win e #fail, comumente utilizadas para expressar aprovação e reprovação, respectivamente, pelos usuários da comunidade brasileira. Para seleção de língua, utilizamos o parâmetro *lang* da API selecionando o valor 'pt'.

Os rótulos de sentimentos estão associados às *hashtags*, de acordo com o sentimento conotado pelas mesmas. Assim, todo texto contendo a *#win* é classificado como possuindo um sentimento positivo, assim como todo aquele contendo *#fail* é classificado como possuindo um sentimento negativo. *Tweets* contendo ambas as *hashtags* foram descartados do conjunto.

Para avaliar o método de classificação de *tweets*, identificamos a polaridade dos termos que conotam sentimentos, usando o método descrito na Seção 4.4, e selecionamos a classe mais frequente no mesmo (ou, equivalentemente, realizamos a soma algébrica das polaridades). Aqueles textos que possuíam mais termos positivos que negativos são, assim, classificados como conotando um sentimento positivo; aqueles com mais termos negativos que positivos conotando sentimento negativo e aqueles com a mesma quantidade de termos positivos e negativos ou nos quais nenhum termo polarizado foi reconhecido, classificou-se como tendo não tendo polaridade ou com polaridade neutra. O resultado da aplicação do método de mineração pode ser visualizado na Tabela 5.3.

Tabela 5.3 – Resultados da classificação de sentimentos sentencial

	Classificação			Métricas		
Anotação	pos	neutro	neg	Prec	Abr	medida F
pos	425	299	126	0,61	0,50	0,55
neg	270	329	251	0,67	0,29	0,40

Note que os resultados estão aquém dos métodos do estado da arte na área - os principais trabalhos citados no Capítulo 3 -, que possuem uma medida F entre 60% e 70% pro caso do Twitter. O principal fator de impacto é ainda a abrangência do método, pois métodos baseados em dicionários são bastante rígidos. Além das deficiências do método em si, percebemos, com a análise dos erros que um agravante nos resultados foi a constituição do corpus.

Ao analisar os *tweets* usados para avaliar o método percebeu-se que 15% dos textos estavam na língua espanhola - como o exemplo abaixo - o que nos fez perceber que o reconhecimento de língua da API do Twitter não é perfeito.

> "no me traje el cargador #fail"

Além disso, muitos dos *tweets* são essencialmente sentenças com informações factuais com meta-informação de opinião expressa pelas *hashtags* #win ou #fail - que foram excluídas do léxico nesse experimento. A opinião é, então, expressa pelas *hashtags* que usamos para construir o corpus. Excluindo-as, as opiniões desses *tweets* poderiam ser classificadas como opiniões implícitas - na terminologia de Liu [50] - um tipo de opinião que não nos propomos a tratar. Um exemplo desse tipo de *tweet* pode ser visto abaixo:

> "comprei um pingo d'ouro ofereci pra todo mundo e ninguém quis #win"

Além disso, descobrimos que cerca de 2% dos tweets do corpus são SPAM que utilizam a *hashtag* #fail como abaixo:

"hora certa 18:10 em brasilia #time #fail #horacerta visite agora nosso chat <LINK>"

5.2.2 Mineração de Opinião sub-sentencial

Para avaliar o método sub-sentencial de Mineração de Opinião, utilizamos o corpus TwitterSentiment - manualmente anotado como discutido acima. Dos 1000 *tweets* anotados do corpus TwitterSentiment, nós selecionamos todas as expressões opinativas anotadas nos textos, totalizando 130 exemplos. Para cada expressão, consultou-se a sua presença no léxico de opiniões OpLexicon. Se o léxico contiver a expressão anotada, as polaridades da anotação e do léxico são comparados, caso contrário, o método classifica a expressão como não-opinativa. O resultado da identificação de polaridade pode ser visualizado na Tabela 5.4.

Tabela 5.4 – Resultado da identificação e classificação de sentimentos

	Método			Métricas		
Anotação	Positivo	Neutro ou Não Opinativo	Negativo	Precisão	Abrangência	medida F
Positivo	17	40	0	0,61	0,30	0,40
Negativo	11	54	8	1	0,12	0,21

Nota-se que, apesar da precisão do nosso método comparável com métodos mais robustos, como o Twittómetro. A abrangência, entretanto, ainda é bastante reduzida. Não é grande surpresa, por tratar-se de um método baseado em dicionários.

Alguns dos problemas observados resultam do escopo limitado das expressões do OpLexicon - em geral constituídas por palavras ou bigramas - e pelo modo de comparação - baseado na presença da expressão completa no léxico. Desse modo, expressões como "rápido e eficiente" não são marcadas como positivas, apesar de ambos os adjetivos da expressão estarem marcados com polaridade positiva no léxico. Um método de composição de opiniões poderia ser utilizado para lidar com essas expressões complexas, outras como "O que seria de nós sem" são, entretanto, mais difíceis e acreditamos que somente através da extensão do léxico poderemos identificar e tratar tais opiniões.

Outro caso muito comum foi o uso de expressões muito particulares de contexto, de região ou do autor como nas expressões "tá dando águia" que expressa uma opinião negativa quanto a entidade "Tweetdeck" na sentença abaixo.

"Tweetdeck tá dando águia aqui no linux :P embaralhando as letras nas colunas :P #fail"

ou ainda, "(x)" expressando uma opinião positiva em relação à entidade "Cristovão Colombo" na sentença

"Qual o melhor navegador? ( ) Firefox. ( ) google Chrome. ( ) Internet Explorer. (x) Cristóvão Colombo."

Acreditamos também que a associação de nosso método com a aplicação de de regras sintáticas - ou pseudo-sintáticas - ou ainda um método de aprendizagem pode melhorar esses resultados, aprendendo contextos indicadores de opinião e a compor tais opiniões parciais na expressão opinativa.

5.3 Reconhecimento de Entidades Nomeadas

Utilizamos para o treino do Reconhecedor de Entidades Nomeadas um conjunto de 900 *tweets* que foi manualmente anotado por somente um anotador, seguindo as mesmas estratégias de anotação para entidades da Seção 5.1. Esses dados foram coletados usando a TwitterAPI buscando *tweets* em língua portuguesa contendo as expressões "google" ou "cielo". O conjunto de textos possui 1459 entidades nomeadas, sendo, dentre elas, 99 constituídas por mais de uma palavra.

Avaliou-se o método discutido no Capítulo 4, inicialmente, utilizando validação cruzada com cinco camadas obteve-se o resultado mostrado na Tabela 5.5.

Tabela 5.5 – Resultados Preliminares do Reconhecimento de Entidades Nomeadas - utilizando o conjunto de desenvolvimento

	Etiquetas					Métricas		
	B	I	L	U	O	Abr	Prec	Medida F
B	22	0	0	6	71	0,22	1,00	0,36
I	0	0	1	0	52	0,00	0,00	0,00
L	0	0	25	0	72	0,26	0,96	0,41
U	0	0	0	392	968	0,29	0,89	0,44
O	0	0	0	41	16801	0,99	0,94	0,96

Note que os baixos índices para a etiqueta "I" (*Inside*) refletem sua infrequência nos dados - apenas 53 casos de EN constituída de 3 ou mais palavras, e.g. "Google Adwords Express".

Devido a anotação ter sido realizada por somente um anotador - sendo assim sujeita a muito viés - e querendo investigar o papel da mudança do domínio dos textos na performance do reconhecedor, utilizamos o TwitterSentiment como conjunto de testes no reconhecedor treinado com os dados de desenvolvimento, discutidos anteriormente. A comparação foi feita da seguinte maneira: a anotação de entidades do TwitterSentiment foi inicialmente convertida para o esquema de anotação BILOU - ignorando as classificação das entidades entre os tipos PESSOA, ORG, PRODUTO e LOC -; os textos - sem a anotação convertida - foram anotados pelo sistema de reconhecimento de entidades treinado com os mesmos dados discutidos na Tabela 5.5; as etiquetas emitidas pelo sistema para cada palavra foi, então, comparada à etiqueta obtida da anotação manual, obtendo os valores mostrados na Tabela 5.6.

Note que o resultado da identificação foi bastante favorável para entidades simples - cujo nome é constituído por somente uma palavra (etiqueta "U"). Nos outros casos, o sistema teve uma performance aquém do esperado - especialmente pelo fato de, no conjunto de desenvolvimento,

Tabela 5.6 – Resultados Finais do Reconhecimento de Entidades Nomeadas - utilizando o recurso dourado

Etiquetas						Métricas		
	B	I	L	U	O	Abr	Prec	Medida F
B	13	0	0	315	212	0,02	0,13	0,03
I	76	0	6	1	63	0,00	0,00	0,00
L	2	0	82	93	359	0,15	0,88	0,26
U	0	0	0	974	323	0,75	0,65	0,70
O	11	0	5	106	20185	0,99	0,95	0,97

os resultados não haverem demonstrado esse baixo desempenho. Um dos possíveis motivos para tanto reside no fato de 2/3 (66%) dos dados do TwitterSentiment tratam de domínio diferente daqueles tratados no conjunto de desenvolvimento - que foi utilizado para treinar o reconhecedor. Evidenciamos assim que, como esperado, a mudança de domínio parece ter impacto significativo no reconhecimento de entidades multi-palavras.

Um problema percebido ao se reconhecer entidades com mais de uma palavra reside, principalmente, nos textos referentes à entidade "Google" e seus produtos (cerca de um terço do corpus). A principal dificuldade para tal caso é que a palavra "google" aparece muito frequentemente como uma entidade unitária - referindo-se comumente à empresa ou à página de busca - ou como o início de uma entidade composta - como em "Google Adsense", "Google Search", "Google Docs", etc - em contextos muito similares. Além disso, essas palavras - "Search", "Docs", etc. - aparecem frequentemente nos textos também como entidades unitárias - e dada a alta ocorrência da palavra "Google" como uma entidade unitária em particular - o reconhecedor tende a classificar tais entidades multipalavras (e.g. "Google Docs") como uma sequência de duas entidades unitárias, ocasionando grande erro.

5.4 Reconhecimento de Referências em Análise de Sentimentos

Para analisar a performance do resolvedor de referência utilizamos o corpus TwitterSentiment, manualmente anotado quanto às Entidades Nomeadas e às opiniões que as referenciam. Fizemos duas avaliações do método: uma considerando a perfeita identificação das entidades e das expressões opinativas, para avaliar o método de resolução sem influência dos erros causados pelos outros dois, e uma seguindo o *framework* estabelecido no capítulo anterior, para analisar a performance do método proposto como um todo.

A anotação de opinião é relativamente infrequente nos textos, dado que em 1000 textos anotados, somente 130 anotações de opinião foram feitas. Com elas, utilizamos validação cruzada em 10 camadas (*10-fold cross-validation*) no método de resolução, obtendo a matriz de confusão ilustrada na Tabela 5.7.

Essa avaliação considera um determinado *tweet* anotado - como abaixo - e avalia se uma dada expressão opinativa refere-se a uma dada entidade - no caso abaixo as expressões "muito massa"

Tabela 5.7 – Matriz de confusão do método de resolução de referência

	Método		Métricas		
Anotação	Referencia	Não Referencia	Prec	Abr	F1
Referencia	94	39	0,69	0,71	0,70
Não Referencia	43	85	0,69	0,66	0,67

que expressa uma opinião positiva aos produtos "Gmail", "Docs" e "Calendar" (*offline*), mas não se refere diretamente à organização "Google". Neste caso, o sistema deve responder verdadeiro quando questionado se a expressão refere-se a uma dessas três primeiras entidades e falso quanto à ultima.

> "RT USUARIO: #NOVIDADE: <ORG>Google</ORG> lança <PRODUTO>Gmail </PRODUTO>, <PRODUTO>Calendar</PRODUTO> e <PRODUTO>Docs </PRODUTO> offline. <LINK> // estou usando desde ontem!!! **muito massa**!!!"

Um atributo que é muito importante no modelo treinado parece ser a distância entre a expressão e sua referência. De fato, muitos dos casos expressão-entidade que o sistema classifica como não relacionados estão separados por uma grande quantidade de caracteres, como no exemplo (1) abaixo. Outros, mais frequentes, nos quais o sistema classifica como relacionados quando a expressão e o candidato a referência estão próximos (2).

> "Extamente. RT @USER: Maldito <PRODUTO>Ubuntu 11.10</PRODUTO>, tentando ser <PRODUTO>Mac OS</PRODUTO> ... Podia ter continuado sendo um bom <PRODUTO>linux</PRODUTO> #fail", Expressão: "Maldito", Candidato:"linux", Resposta: False (1)
>
> "Bah! @USER, a resposta é: não compartilha... Agora é tudo pelo <PRODUTO>Google+</PRODUTO>, um #fail do caramba. Eu era muito fã do <PRODUTO>Google Reader</PRODUTO>", Expressão: "fã", Candidato: "Google+", Resposta: True (2)

De fato, o sistema tende a associar a opinião com a entidade mais próxima em posição anterior à expressão, como é o caso de (2).

O sistema como um todo pode ser avaliado com os resultados da Tabela 5.8.

Tabela 5.8 – Matriz de confusão do método de resolução de referência considerando todos os métodos

	Método		Métricas		
Anotação	Referencia	Não Referencia	Prec	Abr	F1
Referencia	21	8	0,02	0,72	0,04
Não Referencia	1142	984	0,99	0,46	0,63

Note que, apesar da identificação de referência possuir uma medida F bastante baixa, percebe-se que muitos dos erros foram cometidos pela quantidade de termos polarizados encontrados na etapa

de Análise de Sentimentos. Como a quantidade de termos polarizados referindo-se a uma entidade é extremamente mais infrequente que aqueles que não se referem - e no conjunto de treino aqueles que não se referiam a entidade alguma não foram marcados - há uma tendência natural do classificador relacionar mais pares termo-entidade, causando a baixa precisão.

Um exemplo desse tipo de fenômeno pode ser visto abaixo, no qual o autor identifica seu estado privado em relação a um evento, mas não expressa uma avaliação da entidade mencionada no texto:

> "pense me uma pessoa **feliz** pq finalmente consegui instalar o drive do moden no <PRODUTO> linux </PRODUTO> #win"

Perceba ainda que a abrangência do reconhecimento de referências é alta, o que indica que o sistema, em geral, identifica as opiniões verdadeiras. Como a precisão dos casos de não referência é quase perfeita - 0,99 - e levando em consideração que, dentro do processo de inteligência, o analista necessariamente precisará estudar os resultados do método para identificar inteligência, nosso método potencialmente reduz seu trabalho manual, sem esconder informação relevante para a sua análise.

6. Considerações Finais

6.1 Considerações Finais

Este trabalho apresentou um conjunto de métodos visando auxiliar a aplicação processos de Inteligência Competitiva através da aplicação de técnicas de Análise Automática de Texto, Processamento de Linguagem Natural e Extração de Informações. A principal construção desse trabalho é um método sub-sentencial de Análise de Sentimentos focado em entidades para os textos do Twitter a partir das seguintes tarefas: Análise de Sentimentos (AS), Reconhecimento de Entidades Nomeadas (REN) e Reconhecimento de Referência de Opiniões.

Nosso trabalho surge da necessidade de ferramentas de auxílio para análise da resposta do mercado consumidor a um dado produto ou marca explorando o conteúdo das mídias sociais e da constatação que as ferramentas de Análise de Sentimentos atualmente disponíveis não tratam satisfatoriamente o problema da identificação de referência da opinião.

Os resultados obtidos através das avaliações feitas no Capítulo 5, apesar de poderem ainda ser melhorados, considerando o foco do trabalho e o panorama atual das ferramentas para tanto, avaliamos que nossos resultados são satisfatórios e mostram a viabilidade do método aqui proposto. Consideramos ainda que nosso estudo foi, como um todo, satisfatório como um estudo inicial a diferentes problemas ainda inexplorados ou pouco explorados para a língua portuguesa e a quantidade de recursos gerados e de métodos testados para esses problemas.

Atingimos nesse trabalho diversos problemas como a Análise de Sentimentos sentencial e sub-sentencial em dados do Twitter, para o qual atingimos uma performance de cerca de 0,5 e medida F para o caso sentencial e cerca de 0,3 para o caso sub-sentencial - inerentemente mais difícil. Além disso, produzimos um reconhecedor de entidades nomeadas especializado para os textos do Twitter em Português, atingindo uma performance de 0,70 de medida F para entidades de nomes simples. Por fim, desenvolvemos ainda um método de detecção de referências de opiniões a entidades mencionadas em texto - com medida F de 0,70 - além de um léxico e um recurso dourado de análise de sentimentos sub-sentencial com textos provenientes do Twitter para a língua portuguesa.

Nosso trabalho apresenta um avanço às estratégias apresentadas em [37] que utiliza um método sentencial de Análise de Sentimentos e não detecta quais entidades estão sendo mencionadas num *tweet*. Assemelha-se bastante ao trabalho de Silva *et al.* [89] que utiliza padrões lexicossintáticos para detectar opinião (sub-sentencial) no domínio da política e ao de Ding *et al.* que analisa fóruns *online* sobre produtos. Diferentemente do primeiro, o nosso não se baseia em um domínio específico e, diferentemente do trabalho de Ding *et al.*, nós utilizamos o Twitter como fonte - muito mais ruidosa - e analisamos a nível de entidade, não de atributo, o que, cremos, possibilita avanços mais significativos em performance, que um nível tão granular quanto o proposto por esses autores.

Consideramos que a Análise de Texto é uma importante área com vasta aplicação a problemas reais e com resultados práticos para serem aplicados no mercado. Esse trabalho coloca-se como um

exemplo para a aplicação dos métodos que compõem tal área à Análise de Mercado e Inteligência Competitiva.

6.2 Contribuições

Nessa seção relacionamos algumas das contribuições deste trabalho ao contexto acadêmico. São elas:

- Contribuições principais
  - Método de Análise de Sentimentos Subsentencial focado em entidades para o Twitter;
  - Sistema para validação do método mencionado anteriormente - implementado em Python com utilização das bibliotecas NLTK e SciKits;
  - Construção das heurísticas de pré-processamento lexical;
  - Avaliação dos resultados
- Recursos
  - Léxico de Sentimentos independente de domínio para o Português Brasileiro;
  - Diferentes corpora anotado para Análise de Sentimentos e Reconhecimento de Entidades Nomeadas: CineCorpus, TwittEntity e TwitterSentiment;
- Artigos e Trabalhos Relacionados
  - "Construction of a portuguese opinion lexicon from multiple resources," publicado no STIL-2011 em Cuiabá contendo sobre a criação do léxico de sentimentos - correspondendo ao trabalho de final de 2010 e início de 2011.
  - "Sentiment Analysis on Twitter Data for Portuguese Language" aceito para publicação no PROPOR 2012 contendo os resultados dos experimentos de classificação de sentimentos no Twitter, correspondendo ao trabalho do primeiro semestre de 2011
  - Experimentos diversos com membros do laboratório e colaboradores externos que geraram mais 2 artigos publicados com temas relacionados.

6.3 Trabalhos Futuros

Listamos aqui algumas das ideias que surgiram durante o trabalho e que não puderam ser realizadas que, pretende-se, tornem-se trabalhos futuros.

Pretendemos, primeiramente, melhorar o método de Análise de Sentimentos utilizando informação pseudo-sintática e discursiva além de experimentarmos uma estratégia que combine o método baseado em léxico e heurísticas - tais quais a de [20] - e um algoritmo de aprendizagem de máquina.

É interessante também implementar o método de [51] baseado em *co-training* para Reconhecimento de Entidades Nomeadas utilizando uma implementação diferente do classificador CRF para Python. Já que com esse método pode-se especializar o reconhecedor para diversos domínios sem necessariamente precisarmos anotar uma grande quantidade de *tweets* de treino.

Por fim, acreditamos que o método de resolução de referência pode ser ainda mais poderoso se combinado com ontologias de domínio, ao permitir mapear não somente a entidade a que a opinião se refere, mas também quais de seus atributos estão sendo avaliados.

Bibliografia

[1] D. A. Aaker. "Marcas: *Brand Equity* - gerenciando o valor da marca". Negócio, 1998, 310 p.

[2] R. V. X. Aires. "Implementação, adaptação, combinação e avaliação de etiquetidores para o português do brasil". Dissertação de Mestrado, Universidade de São Paulo, 2000, 154 p.

[3] C. Akkaya, J. Wiebe, and R. Mihalcea. "Subjectivity word sense disambiguation". In *14th Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 190–199.

[4] D. G. Altman. "Practical Statistics for Medical Research". Chapman & Hall, 1 edição, 1990.

[5] A. Banfield. "Unspeakable sentences". Boston, EUA: Routledge and Kegan Paul, 1982, 340 p.

[6] L. Barbosa and J. Feng. "Robust sentiment detection on twitter from biased and noisy data". In *23rd International Conference on Computational Linguistics*, 2010, pp. 36–44.

[7] E. Bick. "Functional aspects in portuguese NER". In *7th Workshop on Computational Processing of Written and Spoken Portuguese*, 2006, pp. 80–89.

[8] A. Bifet and E. Frank. "Sentiment knowledge discovery in Twitter streaming data". In *13th International Conference on Discovery Science*, 2010, pp. 1–15.

[9] E. Brill. "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging". *Computational Linguistics*. vol. 21-4, Dez 1995. pp. 543–565.

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

[11] N. Cardoso. "REMBRANDT - reconhecimento de entidades mencionadas baseado em relações e análise detalhada do texto". In *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008, pp. 195–211.

[12] N. Cardoso and D. Santos. "Directivas para a identificação e classificação semântica na colecção dourada do HAREM". In *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007, pp. 211–238.

[13] D. Charlett, R. Garland, and N. Marr. "How damaging is negative word of mouth?". *Marketing Bulletin*. vol. 6-1, 1995. pp. 42–50.

[14] Y. Choi and C. Cardie. "Learning with compositional semantics as structural inference for subsentential sentiment analysis". In *13th Conference on Empirical Methods in Natural Language Processing*, 2008, pp. 793–801.

[15] R. Dale and P. Mazur. "Handling conjunctions in named entities". In *8th International Conference on Computational Linguistics and Intelligent Text Processing*, 2007, pp. 131–142.

[16] D. Davidov, O. Tsur, and A. Rappoport. "Enhanced sentiment learning using twitter hashtags and smileys". In *23rd International Conference on Computational Linguistics*, 2010, pp. 241–249.

[17] N. A. Diakopoulos and D. A. Shamma. "Characterizing debate performance via aggregated twitter sentiment". In *28th international conference on Human factors in computing systems*, 2010, pp. 1195–1198.

[18] X. Ding and B. Liu. "Resolving object and attribute coreference in opinion mining". In *23rd International Conference on Computational Linguistics*, 2010, pp. 268–276.

[19] X. Ding, B. Liu, and P. S. Yu. "A holistic lexicon-based approach to opinion mining". In *1st International Conference on Web search and web data mining*, 2008, pp. 231–240.

[20] X. Ding, B. Liu, and L. Zhang. "Entity discovery and assignment for opinion mining applications". In *15th International Conference on Knowledge Discovery and Data mining*, 2009, pp. 1125–1134.

[21] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel. "The automatic content extraction (ace) program – tasks, data, and evaluation". In *4th International Conference on Language Resources and Evaluation*, 2004, pp. 837–840.

[22] D. Dourado. "Modelos de negócio nas mídias sociais". In *#Mídias Sociais: perspectivas, tendências e reflexões*. PaperCliq, 2010.

[23] A. Esuli and F. Sebastiani. "Determining the semantic orientation of terms through gloss classification". In *14th International Conference on Information and Knowledge Management*, 2005, pp. 617–624.

[24] P. Ferragina and U. Scaiella. "Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)". In *19th ACM international conference on Information and knowledge management*, 2010, pp. 1625–1628.

[25] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. "Annotating named entities in Twitter data with crowdsourcing". In *NAACL Workshop on Creating Speech and Text Language Data With Amazon's Mechanical Turk*, 2010.

[26] L. Fuld. "Inteligência competitiva". Rio de Janeiro: Elsevier-Campus, 2007, 256 p.

[27] A. Go, R. Bhayani, and L. Huang. *Twitter Sentiment Classification using Distant Supervision*. Relatório Técnico, Stanford University, 2009, 6 p.

[28] A. Go, L. Huang, and R. Bhayani. *Twitter Sentiment Analysis*. Relatório Técnico, Stanford University, 17 p.

[29] S. A. Golder and M. W. Macy. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures". *Science*. vol. 333-6051, Set 2011. pp. 1878–1881.

[30] E. Gomes and F. Braga. "Inteligência competitiva". Rio de Janeiro: Ed. Campus, 2002, 140 p.

[31] G. Grefenstette, Y. Qu, J. G. Shanahan, and D. A. Evans. "Coupling niche browsers and affect analysis for an opinion mining application". In *7th International Conference on Computer-Assisted Information Retrieval*, 2004, pp. 186–194.

[32] R. Grishman and B. Sundheim. "Design of the MUC-6 evaluation". In *6th Message Understanding Conference*, 1995, pp. 1–11.

[33] A. Harb, M. Plantié, G. Dray, M. Roche, F. Trousset, and P. Poncelet. "Web opinion mining: how to extract opinions from blogs?". In *5th International Conference on Soft Computing as Transdisciplinary Science and Technology*, 2008, pp. 211–217.

[34] V. Hatzivassiloglou and K. R. McKeown. "Predicting the semantic orientation of adjectives". In *8th Conference on European Chapter of the Association for Computational Linguistics*, 1997, pp. 174–181.

[35] J. P. Herring. "Key intelligence topics: A process to identify and define intelligence needs". *Competitive Intelligence Review*. vol. 10-2, Abr 1999. pp. 4–14.

[36] M. Hu and B. Liu. "Mining and summarizing customer reviews". In *10th International Conference on Knowledge Discovery and Data mining*, 2004, pp. 168–177.

[37] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. "Twitter power: Tweets as electronic word of mouth". *Journal of the American Society for Information Science and Technology*. vol. 60-11, Nov 2009. pp. 2169–2188.

[38] J. Kamps, M. Marx, R. J. Mokken, and M. de Rijke. "Using WordNet to measure semantic orientation of adjectives". In *4th International Conference on Language Resources and Evaluation*, 2004.

[39] A. M. Kaplan and M. Haenlein. "Users of the world, unite! the challenges and opportunities of social media". *Business Horizons*. vol. 53-1, Jan 2010. pp. 59–68.

[40] S.-M. Kim and E. Hovy. "Determining the sentiment of opinions". In *20th International Conference on Computational Linguistics*, 2004.

[41] S.-M. Kim and E. Hovy. "Extracting opinions, opinion holders, and topics expressed in online news media text". In *Workshop on Sentiment and Subjectivity in Text*, 2006, pp. 1–8.

[42] N. Kobayashi, R. Iida, K. Inui, and Y. Matsumoto. "Opinion extraction using a learning-based anaphora resolution technique". In *2nd International Joint Conference on Natural Language Processing*, 2004, pp. 175–180.

[43] N. Kobayashi, K. Inui, and Y. Matsumoto. "Extracting aspect-evaluation and aspect-of relations in opinion mining". In *12th Conference on Empirical Methods in Natural Language Processing*, 2007.

[44] E. Kouloumpis, T. Wilson, and J. Moore. "Twitter sentiment analysis: The good the bad and the omg!". In *5th International AAAI Conference on Weblogs and Social Media*, 2011.

[45] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data". In *8th International Conference on Machine Learning*, 2001, pp. 282–289.

[46] P. Levy. "O que é o virtual". São Paulo: Editora 34, 1996, 160 p.

[47] P. Levy. "A inteligência coletiva: por uma antropologia do ciberespaço". São Paulo: Loyola, 2000, 212 p.

[48] X.-L. Li, L. Zhang, B. Liu, and S.-K. Ng. "Distributional similarity vs. pu learning for entity set expansion". In *48th Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 359–364.

[49] W. Lin, T. Wilson, J. Wiebe, and A. Hauptmann. "Which side are you on? identifying perspectives at the document and sentence levels". In *10th Conference on Natural Language Learning*, 2006, pp. 109–116.

[50] B. Liu. "Sentiment analysis and subjectivity". In *Handbook of Natural Language Processing*. Boca Raton, EUA: CRC Press, Taylor and Francis Group, 2010.

[51] X. Liu, S. Zhang, F. Wei, and M. Zhou. "Recognizing named entities in tweets.". In *ACL*, 2011, pp. 359–367.

[52] B. Locke and J. Martin. *Named Entity Recognition: Adapting to Microblogging*. Relatório Técnico, University of Colorado, 2009.

[53] E. G. Maziero, T. A. S. Pardo, A. Di Felippo, and B. C. Dias-da Silva. "A base de dados lexical e a interface web do TeP 2.0: thesaurus eletrônico para o Português do Brasil". In *14th Brazilian Symposium on Multimedia and the Web*, 2008, pp. 390–392.

[54] R. Mihalcea, C. Banea, and J. Wiebe. "Learning multilingual subjective language via cross-lingual projections". In *45th Annual Meeting of the Association of Computational Linguistics*, 2007, pp. 976–983.

[55] K. Moilanen and S. Pulman. "Sentiment composition". In *International Conference on Recent Advances in Natural Language Processing*, 2007, pp. 378–382.

[56] M. C. Muniz, M. das Gracas V. Nunes, and E. Laporte. "Unitex-pb, a set of flexible language resources for brazilian portuguese". In *Information and Language Technology Workshop*, 2005, pp. 2059–2068.

[57] W. Murnane. "Improving accuracy of named entity recognition on social media data". Dissertação de Mestrado, University of Maryland, 2010.

[58] D. Nadeau. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. Tese de Doutorado, University of Ottawa, 2007.

[59] D. Nadeau and S. Sekine. "A survey of named entity recognition and classification". *Linguisticae Investigationes*. vol. 30-1, January 2007. pp. 3–26.

[60] NIST. *Automatic Content Extraction 2008 Evaluation Plan (ACE08) – Assessment of Detection and Recognition of Entities and Relations within and across Documents*. Relatório Técnico, NIST, 2007, 16 p.

[61] A. C. Oliveira. "Inteligência competitiva na internet". Brasport, 2006, 84 p.

[62] A. Pak and P. Paroubek. "Twitter as a corpus for sentiment analysis and opinion mining". In *7th International Conference on Language Resources and Evaluation*, 2010, pp. 1320–1326.

[63] B. Pang and L. Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts". In *42th Annual Meeting of the Association for Computational Liguistics Conference*, 2004, pp. 271–278.

[64] B. Pang and L. Lee. "Opinion mining and sentiment analysis". *Foundations and Trends® in Information Retrieval*. vol. 2-1–2, 2008. pp. 1–135.

[65] B. Pang, L. Lee, and S. Vaithyanathan. "Thumbs up? Sentiment classification using machine learning techniques". In *7th Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.

[66] A. Passos. "Inteligência competitiva - como fazer ic acontecer na sua empresa". LCTE, 2005, 168 p.

[67] R. Pompéia. "Planejar é mais que preciso". In *#Mídias Sociais: perspectivas, tendências e reflexões*. PaperCliq, 2010.

[68] A.-M. Popescu and O. Etzioni. "Extracting product features and opinions from reviews". In *10th Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 339–346.

[69] M. Porter. "Estratégia competitiva". Rio de Janeiro: Editora Campus, 1986.

[70] L. Ratinov and D. Roth. "Design challenges and misconceptions in named entity recognition". In *13th Conference on Computational Natural Language Learning*, 2009, pp. 147–155.

[71] R. Recuero. "Weblogs, webrings e comunidades virtuais". *Revista 404notfound-Revista Eletrônica do Grupo Ciberpesquisa*. vol. 31, 2003.

[72] R. Recuero. "Teoria das redes e redes sociais na internet: considerações sobre o orkut, os weblogs e os fotologs". 2004.

[73] R. Recuero. "Um estudo do capital social gerado a partir de redes sociais no orkut e nos weblogs". *Revista FAMECOS: mídia, cultura e tecnologia*. vol. 1-28, 2006.

[74] R. Recuero. "Webrings: as redes de sociabilidade e os weblogs". *Sessões do Imaginário-Cinema/ Cibercultura/ Tecnologia da Imagem*. vol. 9-11, 2006.

[75] R. Recuero, R. Araujo, and G. Zago. "How does social capital affect retweets?". *5th International AAAI Conference on Weblogs and Social Media*. 2011.

[76] R. Recuero and G. Zago. "Em busca das 'redes que importam': Redes sociais e capital social no twitter". *Líbero*. vol. 12-24, 2009. pp. 81–94.

[77] R. Recuero and G. Zago. "Rt, por favor": considerações sobre a difusão de informações no twitter". *Revista Fronteiras: estudos midiáticos*. vol. 12-2, 2010.

[78] R. C. Recuero, 2007. "Weblogs, webrings e comunidades virtuais. 2003".

[79] M. L. Richins. "Word-of-mouth communication as negative information". *Advances in Consumer Research*. vol. 11, 1984. pp. 697–702.

[80] E. Riloff and J. Shepherd. "A corpus-based approach for building semantic lexicons". In *2nd Conference on Empirical Methods in Natural Language Processing*, 1997, pp. 117–124.

[81] E. Riloff, J. Wiebe, and W. Phillips. "Exploiting subjectivity classification to improve information extraction". In *20th National Conference on Artificial Intelligence*, 2005, pp. 1106–1111.

[82] A. Ritter, S. Clark, Mausam, and O. Etzioni. "Named entity recognition in tweets: An experimental study". In *16th Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1524–1534.

[83] N. T. Roman. *Emoção e a Sumarização Automática de Diálogos*. Tese de Doutorado, Universidade de Campinas, 2007.

[84] D. Santos. "O modelo semântico usado no primeiro HAREM". In *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007, pp. 43–57.

[85] D. Santos and N. Cardoso. "Breve introdução ao harem". In *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007, pp. 1–16.

[86] S. A. Schwenter. "The pragmatics of negation in brazilian portuguese". *Lingua*. vol. 115-10, Out 2005. pp. 1427–1456.

[87] I. S. Silva, J. Gomide, A. Veloso, W. Meira, Jr., and R. Ferreira. "Effective sentiment stream analysis with self-augmenting training and demand-driven projection". In *34th international ACM SIGIR conference on Research and development in Information*, 2011, pp. 475–484.

[88] M. J. Silva, P. Carvalho, C. Costa, and L. Sarmento. *Automatic Expansion of a Social Judgment Lexicon for Sentiment Analysis*. Relatório Técnico, Universidade of Lisboa, 2010.

[89] M. J. Silva and REACTION TEAM. *Notas sobre a Realização e Qualidade do Twitómetro*. Relatório Técnico, Universidade of Lisboa, 2011.

[90] M. Souza, R. Vieira, D. Busetti, R. Chishman, and I. M. Alves. "Construction of a portuguese opinion lexicon from multiple resources". In *8th Brazilian Symposium in Information and Human Language Technology*, 2011.

[91] V. Stoyanov and C. Cardie. "Toward opinion summarization: Linking the sources". In *Workshop on Sentiment and Subjectivity in Text*, 2006, pp. 9–14.

[92] V. Stoyanov, C. Cardie, and J. Wiebe. "Multi-perspective question answering using the OpQA corpus". In *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 923–930.

[93] F. Su and K. Markert. "From words to senses: a case study of subjectivity recognition". In *22th International Conference on Computational Linguistics*, 2008, pp. 825–832.

[94] B. Sundheim. "Named entity task definition (version 2.1)". In *6th Message Understanding Conference*, 1995.

[95] E. F. Tjong Kim Sang and F. De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition". In *Proceedings of CoNLL-2003*, 2003, pp. 142–147.

[96] P. D. Turney. "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews". In *40th Annual Meeting on Association for Computational Linguistics*, 2002, pp. 417–424.

[97] J. Wiebe. "Tracking point of view in narrative". *Computational Linguistics*. vol. 20-2, Jun 1994. pp. 233–287.

[98] J. Wiebe and R. Bruce. "Probabilistic classifiers for tracking point of view". In *Symposium on Empirical Methods in Discourse Interpretation and Generation*, 1995, pp. 181–187.

[99] M. Wiegand, A. Balahur, B. Roth, D. Klakow, and A. Montoyo. "A survey on the role of negation in sentiment analysis". In *Workshop on Negation and Speculation in Natural Language Processing*, 2010, pp. 60–68.

[100] T. Wilson, J. Wiebe, and P. Hoffmann. "Recognizing contextual polarity in phrase-level sentiment analysis.". In *10th Conference on Empirical Methods in Natural Language Processing*, 2005.

[101] T. Wilson, J. Wiebe, and R. Hwa. "Recognizing strong and weak opinion clauses". *Computational Intelligence*. vol. 22-2, Mai 2006. pp. 73–99.

[102] T. A. Wilson. *Fine-Grained Subjectivity and Sentiment Analysis: Recognizing the Intensity, Polarity, and Attitudes of Private States*. Tese de Doutorado, University of Pittsburg, 2008.

[103] Y. Wu, Q. Zhang, X. Huang, and L. Wu. "Phrase dependency parsing for opinion mining". In *14th Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 1533–1541.

Apêndice A. Instruções fornecidas para os anotadores

A primeira parte é a identificação das Entidades Nomeadas. Entidades Nomeadas, i.e. entidades mencionadas em textos através de designadores rígidos como nomes próprios, expressões temporais e espécies biológicas - nomes que não mudam de acordo com a referência temporal ou espacial como "a secretária", etc.

Instruções para anotação das entidades:

* Serão anotadas somente entidades pertencentes às classes Pessoa (<PESSOA>) - i.e. um indivíduo ou grupo de pessoas referido por um nome; Organização (<ORG>)- entidades organizacionais como empresas, organizações, partidos, etc. referidos por um nome; Local (<LOC>) - entidade geopolítica que possui um nome e Produto (<PRODUTO>)- objeto que tem um nome individualizado ou designa elementos que não têm nome individual mas que são designados pelo nome da classe a que pertencem, tal como Ford ou iPod em o meu Ford ou o Ipod dela; * Se houver mais de uma opção de se anotar um determinado trecho textual, será priorizada a anotação que maximize o tamanho dos nomes, e.g. no trecho "Samsung Galaxy S" a anotação "<PRODUTO>Samsung Galaxy S</PRODUTO>" é priorizada sobre "<ORG>Samsung</ORG> <PRODUTO>Galaxy S</PRODUTO>", assim como no trecho "Corpo de Bombeiros de Porto Alegre", a anotação "<ORG>Corpo de bombeiros de Porto Alegre</ORG>" é priorizada sobre "<ORG>Corpo de Bombeiros</ORG> de <LOC>Porto Alegre</LOC>".

A segunda parte consiste de identificar a opinião em textos.

Um segmento textual pode apresentar conteúdos factuais - ou seja, apresentar fatos - ou subjetivos - ou seja, conteúdo que expressa o estado privado do autor, seja suas opiniões, crenças etc. Nesse experimento procuramos anotar os estados privados avaliativos, i.e. os estados privados em que um usuário expressa um sentimento ou opinião quanto a uma entidade explicitamente. Como no exemplo:

"ATÉ QUE ENFIM chegou maquina nova da <ORG>Cielo</ORG>, oooo empresinha enrolada."

Nesse contexto, "Cielo"nomeia uma entidade do tipo organização que está sendo avaliada através da conteúdo textual "enrolada"pelo autor. A expressão "enrolada"conota um sentimento negativo que o autor possui em relação à entidade "Cielo". Assim tal texto seria anotado como:

"ATÉ QUE ENFIM chegou maquina nova da <ORG>Cielo</ORG>, oooo empresinha [enrolada, neg, Cielo]."

Uma expressão avaliativa, i.e. um trecho textual que conota um sentimento pode ser constituído de uma unidade lexical, como no caso acima, ou de uma expressão multipalavras como abaixo:

"Acabei de mexer no <PRODUTO>Xoom</PRODUTO>. [Não achei sequer uma desvantagem,pos,Xoom] em relação ao <PRODUTO>iPad</PRODUTO>! [Muito bom,pos,Xoom]!"

Serão contextos avaliativos que refiram-se somente a entidades. No caso abaixo,

"consegui usar máquina da <ORG>cielo</ORG> sozinha e passei a 2 compra da cliente sem

sustos. FUCK YEAH"

como a expressão avaliativa "FUCK YEAH"refere-se ao evento e não à entidade mencionada no contexto, não dever anotada.