2
Sentiment Analysis

2.1
Definição do Problema

Sentiment Analysis é um problema de categorização de texto no qual deseja-se detectar opiniões favoráveis e desfavoráveis com relação a um determinado tópico (e.g. organizações e seus produtos).

Nos últimos anos, é crescente o interesse da comunidade por *sentiment analysis* onde documentos são classificados pelo sentimento, conotação, atitudes e opiniões ao invés de se restringir aos fatos descritos neste.

O principal desafio em *sentiment analysis* é identificar como sentimentos são expressados em textos e se tais sentimentos indicam uma opinião positiva (favorável) ou negativa (desfavorável) com relação a um tópico.

Uma aplicação do problema é inteligência competitiva, onde deseja-se monitorar o que vem sendo publicado na mídia a respeito de concorrentes, como o público reagiu ao lançamento de um produto, ou até mesmo como a imagem da empresa vem evoluindo ao longo das últimas semanas.

Tradicionalmente, os setores de marketing realizam pesquisas de mercado com este propósito. Embora Pesquisa de Mercado seja uma área bastante madura e quando bem elaboradas gerem boas estimativas, tendem a ser muito custosas principalmente quando trata-se de grandes volumes de dados (06). Com a explosão de informação disponível na Web num ambiente onde todos tendem a ser geradores de conteúdo e expressarem opiniões sobre os mais variados assuntos, aplicações que consolidam opiniões e geram estatísticas relevantes passam a ter um valor significativo.

Obviamente, não faz sentido falar em *sentiment analysis* na Web se não soubermos *onde* estão as opiniões. *Subjectivity Mining* é um problema intimamente relacionado à *Sentiment Analysis* e consiste justamente em *separar* trechos de texto que relatam fatos dos trechos de texto que emitem opiniões.

Porém, apesar das técnicas de *subjectivity mining* nos auxiliarem em identificar as entradas para um classificador de sentimento, estas não en-

dereçam a tarefa específica de identificar a polaridade de uma opinião.

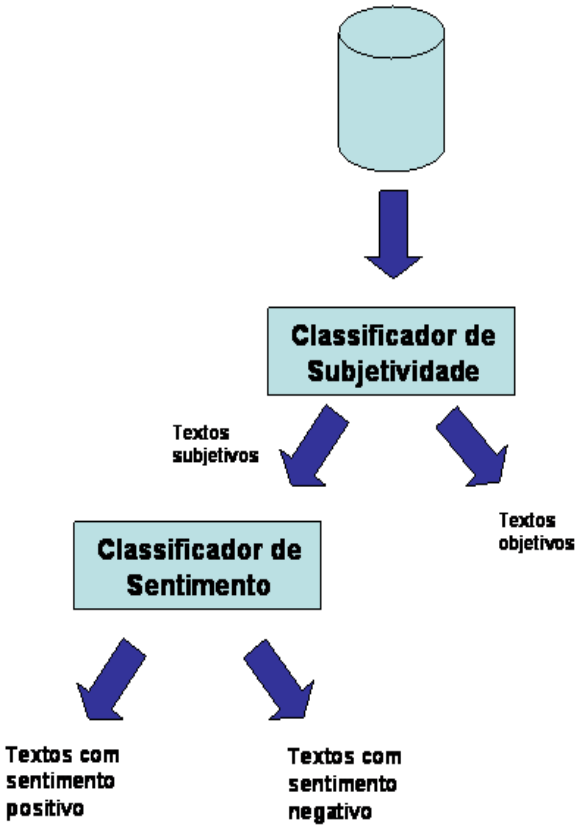

Figura 2.1: Processo de Classificação de Sentimento

2.2 Trabalhos Relacionados

Duas principais abordagens vem sendo aplicadas a *Sentiment Analysis*. A primeira consiste em modelar documentos como sacos-de-palavras e aplicar métodos estatísticos que utilizam as freqüências das palavras para treinar classificadores (19, 13). Esta abordagem baseia-se na teoria estatística de que é possível aprender estruturas complexas de linguagem mesmo que características básicas como a ordem com que as palavras ocorrem sejam ignoradas.

A outra abordagem consiste em classificar textos baseado no *sentiment orientation* de certos termos extraídos através de heurísticas lingüísticas ou de um conjunto de palavras sementes pré-selecionadas (24, 09). Turney apresenta em (24) um algoritmo não supervisionado baseado na *orientação semântica* de bigramas adjetivais presentes nos textos. Neste trabalho, reporta-se uma precisão média de 74% ao utilizar como conjunto de

testes 410 avaliações extraídas do *Epinion*[1] e utilizar dados amostrais de 4 diferentes domínios (avaliações de automóveis, bancos, filmes e destinos de viagem). A precisão varia de 84% para o domínio de automóveis a 66% para avaliações de filmes sugerindo que este último domínio seja mais desafiador. Turney define a palavra "excelent" como sendo referência positiva e a palavra "poor" como referência negativa. A orientação semântica dos bigramas adjetivais é medida como uma relação entre a quantidade de vezes que o bigrama aparece próximo a estas palavras. Esta técnica é considerada não-supervisionada uma vez que não depende de dados rotulados.

O *benchmark* mais usado para avaliar *Sentiment Classifiers* é o *movie review data set* disponibilizado por Pang et al. (19). Este *corpus* foi extraído do IMDb[2] e é composto por 2000 avaliações de filmes (1000 positivas e 1000 negativas). Este *corpus* é descrito detalhadamente no Capítulo 6. Pang compara o desempenho de 3 técnicas de Aprendizado de Máquina neste conjunto de dados: Naive Bayes, Máxima Entropia e Máquinas de Vetores Suporte (SVM). Nos experimentos, é observado que o classificador Naive Bayes tende a ter o pior desempenho e o classificador SVM o melhor. É constatado também que modelos binários (i.e. aqueles que representam as presenças de características) resultam num melhor desempenho do que os modelos que contabilizam as freqüências. Tal fato é um indicador de que classificação de sentimento requer um tratamento diferenciado de categorização de tópicos. A Tabela (2.1) descreve os resultados reportados em (19).

O principal objetivo desta dissertação é examinar se é suficiente tratar classificação de sentimento como um caso especial de categorização de tópico (considerando que os dois tópicos são *sentimento positivo* e *sentimento negativo*) ou se métodos específicos para o problema precisam ser desenvolvidos.

Adotamos como principais referências (19) e (17) e procuramos reproduzir os principais experimentos reportados nestes trabalhos.

Features	#features	NB	ME	SVM
unigrams(frequency)	16162	79.0	n/a	73.0
unigrams	16162	81.0	80.2	82.9
unigrams+bigrams	32324	80.7	80.7	82.8
bigrams	16162	77.3	77.5	76.5
unigrams+POS	16668	81.3	80.3	82.0
adjetivos	2631	76.6	77.6	75.3
top 2621 unigrams	2631	80.9	81.3	81.2
unigrmas+position	22407	80.8	79.8	81.8

Tabela 2.1: Resultados reportados em Pang et al. (19)

[1] www.epinion.com
[2] www.imdb.com.br

Na Tabela (2.1), *NB* representa o classicador Naive Bayes, *ME* o classificador de Máxima Entropia e *SVM* o classificador Máquinas de Vetores Suporte. *unigrams+pos* representa o conjunto de dados em que as palavras são rotuladas com suas classes gramaticais (POS = *Part of Speech*). *unigrmas+position* é uma representação do documento que leva em consideração o *trecho* do texto em que a palavra ocorre.

O problema de *Subjectivity Filtering* é abordado em (17) através de um algoritmo de corte mínimo em grafos. Pang reporta uma melhoria de 82.8% para 86.4% no classificador Naive Bayes porém a técnica não surte efeito no classificador SVM.

Os melhores resultados até agora obtidos usando o *movie review data set* são reportados por Matsumoto et al (12). Neste trabalho, os autores apresentam modelagens de *features* complexas como subseqüência de palavras e árvore de dependência. Neste modelo, todas as subseqüências de uma frase são modeladas como uma característica do texto. Por exemplo, a simples frase de 4 palavras "*I enjoyed the movie*" gera as 15 *features* listadas a seguir:

- I
- I enjoyed
- I the
- I movie
- I enjoyed the
- I enjoyed movie
- I the movie
- I enjoyed the movie
- enjoyed
- enjoyed the
- enjoyed movie
- enjoyed the movie
- the
- the movie
- movie

As árvores de dependência semântica entre as palavras são geradas usando um pacote lingüístico. Este trabalho reporta uma precisão de 93.7% sendo este o melhor resultado publicado para *movie review data set*. Porém,

questionamos a escalabilidade deste modelo uma vez que o número de *features* é exponencial em função do número de palavras nas frases.

Whitelaw et al. (26) buscam melhorar o desempenho do classificadores através de uma análise taxonômica das frases e definindo elementos que compõe uma opinião. Neste trabalho demonstram que estas técnicas em conjunto com sacos-de-palavras resulta numa precisão de 90%.