1
Introdução

Organizar o conhecimento humano em classes relacionadas é quase tão antigo quanto o próprio conhecimento humano. Durante muitos séculos documentos foram classificados manualmente e até recentemente grandes acervos eram organizados desta forma.

O desenvolvimento de métodos de aprendizado estatísticos impulsionados pelo crescente poder computacional disponível e o fácil acesso à massivos volumes de dados rotulados tem tornado técnicas de aprendizado de máquina cada vez mais atraentes.

A construção de máquinas capazes de aprender a partir de experiência tem sido objeto de discussão filosófica e técnica por muito tempo. Hoje, sabemos que máquinas podem receber um nível significativo de capacidade de aprendizado apesar de os limites de tal capacidade estarem ainda muito pouco definidos.

1.1
Objetivos da Dissertação

Consideramos o problema de aprendizado supervisionado onde dados D exemplos de treinamento $\{x_i, y_i\}$, $i = 1, 2, ... D$, sendo que x_i representa um documento que expressa uma opinião e $y_i \in \{-1, 1\}$ sua classe correspondente (positiva ou negativa). O objetivo é aprender a mapear $x \mapsto y$ de modo a minimizar o risco de erro, assumindo que x segue uma distribuição similar aos elementos de sua classe.

Analisamos e comparamos algumas técnicas de aprendizado supervisionado para o problema de *Sentiment Classification*. Adotamos como principais referências os modelos apresentados por Pang et al. (19, 17) e procuramos reproduzir os principais experimentos reportados nestes trabalhos. Apresentamos alguns modelos textuais como saco-de-palavras, N-gramas e seleção de características baseado na Informação Mútua Média (IMM). Comparamos o desempenho dos classificadores Naive Bayes e Support Vector Machines (SVM) para cada modelo proposto. Apresentamos também a importância do filtro de subjetividade no desempenho dos classificadores de sentimento.

1.2 Organização da Tese

No capítulo 2, descrevemos o problema de *Sentiment Analysis* e apresentamos uma revisão das principais abordagens adotadas para o problema.

No capítulo 3, apresentamos os modelos textuais usados nos experimentos, isto é, quais características dos documentos são levadas em consideração pelos algoritmos de classificação.

Nos capítulos 4 e 5, apresentamos as famílias de classificadores Naive Bayes e SVM respectivamente.

No capítulo 6, descrevemos o *corpus* e a metodologia de teste usada nos experimentos e discutimos os resultados obtidos com os classificadores.

Finalmente, no Capítulo 7, apresentamos as conclusões deste trabalho.