1
Introdução

Com 2 bilhões de usuários (MINIWATTS MARKETING GROUP, 2011) e mais de 120 milhões de websites (DOMAINTOOLS.COM, 2011), a Internet já alcança 28,7% da população mundial. Para atrair cada vez mais usuários, dentre outras estratégias, os sites estão se tornando mais participativos. Isto pode ser conseguido, principalmente, com integração às redes sociais, incentivo à contribuição de conteúdo, fóruns, listas de discussão ou comentários. Porém com a enorme quantidade de usuários interagindo, torna-se muito difícil controlar o conteúdo gerado. É provável ocorrer alguns excessos, como o uso indevido para disseminação de spams, vírus, pornografia e outras mensagens indesejadas.

Para alguns sites, a própria comunidade de usuários faz a moderação, através de funcionalidades como denuncie e mecanismos de reputação, dentre outros. O problema desta solução é que ela é pós-moderada, ou seja, a mensagem é sempre exibida e, de acordo com a classificação dos usuários, ela poderá ser removida. Porém, até que isto ocorra, vários usuários já a leram.

Entretanto, existem sites em que uma única mensagem indesejada, mesmo que entre muitas, não é sequer admissível. Imagine um comentário pornográfico em um site construído para crianças. Certamente, os mecanismos de pós-moderação não serviriam para este caso. Para alguns sites é necessário que a mensagem seja pré-moderada, ou seja, antes que qualquer usuário do site possa vê-la, um moderador, autorizado pelo site, vai validá-la e aprovar sua publicação, caso o conteúdo seja compatível. Porém, para sites com grande volume de acessos, a quantidade de moderadores necessários para fazer toda a moderação pode inviabilizar o negócio ou muitas contribuições serão deixadas de lado, desmotivando os usuários de participar. Para resolver este problema, o objetivo da dissertação será construir um sistema de moderação de comentários automático, capaz de aprender novos padrões automaticamente através de um moderador humano e auxiliá-lo na tarefa de moderação.

A dificuldade para a realização deste trabalho está na identificação do

critério do que deve ser aprovado ou reprovado, pois normalmente ele é subjetivo. Às vezes é necessário um entendimento do contexto para avaliar se um determinado comentário poderá ser aprovado ou não. Por exemplo, o comentário feito através do Twitter na página da novela Passione durante a campanha eleitoral:

Lá vem o Serra no comercial de #passione

Este comentário foi bloqueado apesar de não possuir nenhuma palavra obscena, pois se tratava de um comentário sobre política e certamente a emissora de TV não iria permitir comentários políticos na página da novela, especialmente durante a campanha, pois precisa manter a imparcialidade.

Apesar da palavra “Serra" ser um indicativo de que o comentário devesse ser reprovado no exemplo acima, o mesmo não deveria ocorrer se o comentário aparecesse em uma página de debate político. Ou seja, é necessário ainda que o sistema leve em consideração o contexto que o comentário aparece para ser mais preciso na classificação. Porém, não é objetivo do trabalho verificar se o comentário pertence ao contexto em que foi inserido, somente ser especializado ao contexto no momento da moderação.

1.1.Organização da tese

Esse trabalho está organizado da seguinte forma. O Capítulo 1 descreve os trabalhos relacionados a este. O Capítulo 3 apresenta a modelagem adotada, incluindo a parte referente aos algoritmos para auto-moderação. O Capítulo 4 mostra a arquitetura do sistema. O Capítulo 5 expõe alguns experimentos feitos com dois corpora reais e os resultados obtidos. O Capítulo 6 resume os resultados obtidos e sugere trabalhos futuros.