1
Introdução

No momento de tomar uma decisão importante, é sempre recomendado consultar a opinião de outras pessoas. Antes da Internet, muitos de nós solicitávamos aos amigos sugestões para comprar um bom carro, bem como indicações de candidatos para as próximas eleições. Com o avanço da Web, passou a ser possível obter informações sobre opiniões e experiências de pessoas completamente desconhecidas. Além do conteúdo disponibilizado por jornais, os usuários de Internet passaram a compartilhar conhecimentos, críticas e opiniões em blogs pessoais, sites de relacionamento, dentre outros meios. Para pesquisadores (01, 12, 16), essa crescente quantidade de dados oferece uma oportunidade de conhecer a consciência coletiva. Essa nova área de pesquisa conhecida como Análise de Sentimentos tenta identificar o sentimento que os usuários apresentam a respeito de alguma entidade de interesse, como, por exemplo, um produto específico, uma empresa, um lugar ou uma pessoa. O levantamento desses sentimentos é usualmente baseado no conteúdo disponível na Web. O objetivo principal é sumarizar automaticamente a opinião coletiva sobre algum item, sem precisar encontrar e ler todas as opiniões e notícias a esse respeito.

Existem diversas aplicações para a Análise de Sentimento (01, 12, 04). A opinião de usuários sobre um determinado produto é uma informação valiosa para a empresa, pois pode ser usada para melhorar o produto ou direcionar ações de marketing. Eleitores podem identificar a opinião de outros eleitores sobre um determinado candidato. Também é possível analisar a opinião de pessoas sobre filmes, jogos e livros. Pesquisas recentes (22, 16, 18, 07, 09) mostram que notícias do mercado financeiro têm uma grande relação com as variáveis de mercado como volume de transações, volatilidade e preço das ações. Uma vez que diversos investidores têm acesso a notícias online, rumores e escândalos podem influenciar o preço das ações, por isso acredita-se que notícias podem conter grandes indicadores e são potencialmente fonte de preciosas informações. Porém, processar todas as informações disponíveis em tempo real é muito complexo tanto para humanos quanto para máquinas.

Neste trabalho, investigamos o problema de análise de sentimentos de

PUC-Rio - Certificação Digital Nº 0912855/CA

notícias jornalísticas sobre o mercado financeiro em português. Nosso objetivo é classificá-las como favoráveis ou não com relação à Petrobras. Existem dois tipos de informação sobre o mercado financeiro, a saber: artigos jornalísticos e textos gerados por usuários da internet. Textos jornalísticos são mais estáveis, não contém gírias ou abreviações de escrita, e por isso são mais confiáveis. Porém, de acordo com Schumaker (16), mesmo quando as informações de jornais aparentam causar um visível impacto no preço de ações, compras ou vendas inesperadas podem mudar a direção do mercado. Como esses eventos são mais raros, nosso foco é em notícias jornalísticas.

Como não existe um corpus para a tarefa de classificação de sentimento de notícias de mercado em português, criamos o PETRONEWS, um corpus anotado em português contendo notícias sobre a Petrobras. A primeira etapa de nossa abordagem é filtrar as frases que estão relacionadas com a Petrobras em cada uma das notícias, ou seja, utilizar apenas frases *no tópico*. Em seguida, criamos alguns novos atributos estruturais e sintáticos, sendo mantidos apenas os mais informativos segundo quatro critérios diferentes de seleção de atributos. Finalmente, construímos um comitê com os melhores modelos SVM.

Modelo	Acurácia (%)
Filtro de tópico e comitê de modelos	87, 14
Filtro de tópico e dicionário de palavras	85, 81
Dicionário de palavras	82, 76

Tabela 1.1: Acurácia dos modelos para o PETRONEWS.

Nossos resultados indicam que filtrar as frases no tópico reduz o erro em 17, 7%. Acrescentar novos atributos estruturais e sintáticos e fazer uma seleção deles reduz o erro em 25, 4% quando comparado ao modelo base de saco-de-palavras.

Esse trabalho está organizado nos seguintes capítulos. No capítulo 2, descrevemos trabalhos sobre a tarefa de análise de sentimentos e predição de mercado financeiro baseada em notícias on-line. No capítulo 3, descrevemos o PETRONEWS, o corpus anotado contendo notícias sobre a Petrobras. No capítulo 4 explicamos o algoritmo de aprendizado de máquina utilizado, o SVM. No capítulo 5 detalhamos como é feito o filtro por tópico, selecionando apenas as frases sobre a Petrobras. No capítulo 6, descrevemos cada um dos atributos utilizados nos experimentos: atributos sintáticos, semânticos e estruturais. No capítulo 7, detalhamos cada um dos métodos de seleção de atributos. No capítulo 8, descrevemos os experimentos. Em seguida no capítulo 9 apresentamos os resultados dos melhores experimentos. Finalmente no capítulo 10 nossas conclusões são apresentadas.

PUC-Rio - Certificação Digital Nº 0912855/CA

No apêndice 1 e 2, descrevemos o conjunto de etiquetas morfossintáticas e de chunks, respectivamente, utilizadas. No apêndice 3, listamos todas as pontuações e caracteres especiais utilizados como atributos estruturais e no apêndice 4, as palavras funcionais.

PUC-Rio - Certificação Digital Nº 0912855/CA