



Paula de Castro Sonnenfeld Vilela

**Classificação de Sentimento para Notícias
sobre a Petrobras no Mercado Financeiro**

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática da PUC-Rio

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro
Julho de 2011



Paula de Castro Sonnenfeld Vilela

**Classificação de Sentimento para Notícias
sobre a Petrobras no Mercado Financeiro**

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Ruy Luiz Milidiú

Orientador

Departamento de Informática — PUC-Rio

Prof. Marco Antonio Casanova

Departamento de Informática — PUC-Rio

Prof. Karin Koogan Breitman

Departamento de Informática — PUC-Rio

Prof. José Eugenio Leal

Coordenador Setorial do Centro Técnico Científico — PUC-Rio

Rio de Janeiro, 01 de Julho de 2011

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

Paula de Castro Sonnenfeld Vilela

Graduou-se em Engenharia Civil com ênfase em estruturas na Pontifícia Universidade Católica do Rio de Janeiro, cursando matérias do departamento de informática como Banco de Dados e Programação Orientada a Objetos. Teve como trabalho de fim de curso o desenvolvimento de um software para calcular concreto protendido. Trabalhou na empresa ZAP Sistemas no desenvolvimento de softwares voltados para o mercado financeiro. Desenvolveu junto com o seus orientadores durante o Mestrado ferramentas de classificação de texto utilizando aprendizado de máquina.

Ficha Catalográfica

Vilela, Paula de Castro Sonnenfeld

Classificação de Sentimento para Notícias sobre a Petrobras no Mercado Financeiro / Paula de Castro Sonnenfeld Vilela; orientador: Ruy Luiz Milidiú. — 2011.

v., 50 f: il. ; 30 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2011.

Inclui referências bibliográficas.

1. Informática – Tese. 2. Análise de Sentimento. 3. Aprendizado de Máquina. 4. SVM. 5. Mineração de opiniões. 6. Processamento de Linguagem Natural. 7. Seleção de Atributos. 8. Classificação de texto. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

Ao meu orientador Professor Ruy Luiz Milidiú por dividir um pouco de seus conhecimentos comigo, pelo seu incentivo para a realização deste trabalho e por me ajudar a conciliar minha vida profissional com a acadêmica.

Ao CNPq e à PUC-Rio, pelos auxílios concedidos, sem os quais este trabalho não poderia ter sido realizado.

Aos meus pais e irmãs por seu apoio incondicional em todos os momentos da minha vida.

Ao meu noivo Daniel pela paciência e compreensão nos últimos anos.

Ao meu amigo Rodrigo por me introduzir a esse mundo da informática e aos meus colegas da PUC-Rio, quem me fizeram adorar esse lugar.

Ao pessoal do departamento de Informática pela ajuda de todos os dias.

Resumo

Vilela, Paula de Castro Sonnenfeld; Milidiú, Ruy Luiz. **Classificação de Sentimento para Notícias sobre a Petrobras no Mercado Financeiro**. Rio de Janeiro, 2011. 50p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Hoje em dia, encontramos uma grande quantidade de informações na internet, em particular, notícias sobre o mercado financeiro. Diversas pesquisas mostram que notícias sobre o mercado financeiro possuem uma grande relação com variáveis de mercado como volume de transações, volatilidade e preço das ações. Nesse trabalho, investigamos o problema de Análise de Sentimentos de notícias jornalísticas do mercado financeiro. Nosso objetivo é classificar notícias como favoráveis ou não a Petrobras. Utilizamos técnicas de Processamento de Linguagem Natural para melhorar a acurácia do modelo clássico de saco-de-palavras. Filtramos frases sobre a Petrobras e inserimos novos atributos linguísticos, tanto sintáticos como estilísticos. Para a classificação do sentimento é utilizado o algoritmo de aprendizado *Support Vector Machine*, sendo aplicados ainda quatro seletores de atributos e um comitê dos melhores modelos. Apresentamos aqui o PETRONEWS, um corpus com notícias em português sobre a Petrobras, anotado manualmente com a informação de sentimento. Esse corpus é composto de mil e cinquenta notícias online de 02/06/2006 a 29/01/2010. Nossos experimentos mostram uma melhora de 5.29% com relação ao modelo saco-de-palavras, atingindo uma acurácia de 87.14%.

Palavras-chave

Análise de Sentimento; Aprendizado de Máquina; SVM; Mineração de opiniões; Processamento de Linguagem Natural; Seleção de Atributos; Classificação de texto;

Abstract

Vilela, Paula de Castro Sonnenfeld; Milidiú, Ruy Luiz(Advisor). **Sentiment Analysis for Financial News about Petrobras Company**. Rio de Janeiro, 2011. 50p. MSc Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A huge amount of information is available online, in particular regarding financial news. Current research indicate that stock news have a strong correlation to market variables such as trade volumes, volatility, stock prices and firm earnings. Here, we investigate a Sentiment Analysis problem for financial news. Our goal is to classify financial news as favorable or unfavorable to Petrobras, an oil and gas company with stocks in the Stock Exchange market. We explore Natural Language Processing techniques in a way to improve the sentiment classification accuracy of a classical bag of words approach. We filter on topic phrases for each Petrobras related news and build syntactic and stylistic input features. For sentiment classification, Support Vector Machines algorithm is used. Moreover we apply four feature selection methods and build a committee of SVM models. Additionally, we introduce PETRONEWS, a Portuguese financial news annotated corpus about Petrobras. It is composed by a collection of one thousand and fifty online financial news from 06/02/2006 to 01/29/2010. Our experiments indicate that our method is 5.29% better than a standard bag-of-words approach, reaching 87.14% accuracy rate for this domain.

Keywords

Sentiment Analysis; Machine Learning; SVM; Natural Language Processing - NLP; Opinion Mining; feature selection; text classification;

Sumário

1	Introdução	11
2	Pesquisas Anteriores	14
3	PETRONEWS	16
3.1	Geração	16
3.2	Características	17
4	Support Vector Machines	19
5	Filtragem Por Tópico	21
6	Atributos	22
6.1	Atributos N-Gramas	23
6.2	Processamento de Linguagem Natural	23
6.3	Atributos Estruturais	25
6.4	Caminhamento na árvore de decisão	25
7	Seleção de Atributos	28
7.1	Proportional Difference - PD	28
7.2	Fisher Score	30
7.3	Entropy Weighted Genetic Algorithm - EWGA	30
7.4	Árvore de decisão	34
8	Experimentos	36
8.1	Impacto do uso de Filtro por tópico	36
8.2	Impacto do uso de atributos estruturais	37
8.3	Impacto do uso de n-gramas de palavras	37
8.4	Impacto do uso de NLP	38
8.5	Impacto do uso de n-gramas de etiquetas Morfossintáticas	38
8.6	Impacto do uso do caminho da árvore de decisão como atributo	39
8.7	Impacto do uso de Seleção de Atributos	39
9	Resultados	41
9.1	Melhores Resultados	41
9.2	Resultados por Categoria	42
10	Conclusão	43
	Referências Bibliográficas	45
A	Conjunto de Etiquetas Morfossintáticas	47
B	Conjunto de Etiquetas Chunks	48
C	Descrição dos Atributos Estruturais - Caracteres Especiais e Pontuação	49

Lista de figuras

4.1	Hiperplano com a Máxima Distância	19
6.1	Exemplo de Árvore de decisão	26
7.1	Exemplo de Árvore de decisão	35

Lista de tabelas

1.1	Acurácia dos modelos para o PETRONEWS.	12
3.1	Distribuição de Notícias do PETRONEWS por Sentimento.	18
4.1	Vetores de entrada para o SVM	20
6.1	Novos atributos gerados pela Árvore de Decisão	27
7.1	Redução do número de atributos e acurácia com uso do seletor PD	29
7.2	Redução do número de atributos com uso do seletor	30
7.3	Atributos selecionados pela Árvore de Decisão	34
8.1	Impacto do uso de Filtro por tópico	36
8.2	Impacto do uso de atributos estruturais	37
8.3	Impacto do uso de n-gramas de palavras	37
8.4	Impacto do uso de NLP	38
8.5	Impacto do uso de n-gramas de etiquetas morfossintáticas	39
8.6	Impacto do uso do caminho na árvore de decisão como atributo	39
8.7	Impacto do uso de Seleção de Atributos	40
9.1	Melhores Resultados	41
9.2	Erro por categoria de notícias	42
A.1	ETIQUETAS MORFOSSINTÁTICAS	47
B.1	ETIQUETAS DE CHUNKS	48
C.1	CARACTERES ESPECIAIS E PONTUAÇÃO	49
D.1	PALAVRAS FUNCIONAIS	50