Pontifícia Universidade Católica
do Rio de Janeiro

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

Pontifícia Universidade Católica
do Rio de Janeiro

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

Paula de Castro Sonnenfeld Vilela

Graduou–se em Engenharia Civil com ênfase em estruturas na Pontifícia Universidade Católida do Rio de Janeiro, cursando matérias do departamento de infomática 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 compreenção 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

Lista de figuras

Lista de tabelas