PHILIPPE SANTA MARIA NIZER

MÉTODO PARA DETECÇÃO DO EFEITO DE PUBLICAÇÕES DE NOTÍCIAS NO MERCADO DE AÇÕES DO BRASIL

CURITIBA

2011

PHILIPPE SANTA MARIA NIZER

MÉTODO PARA DETECÇÃO DO EFEITO DE PUBLICAÇÕES DE NOTÍCIAS NO MERCADO DE AÇÕES DO BRASIL

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para a obtenção do título de Mestre em Informática.

Área de Concentração: *Ciência da Computação.*

Orientador: Prof. Dr. Júlio César Nievola

CURITIBA

2011

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR
Biblioteca Central

Nizer, Philippe Santa Maria
N737m Método para detecção do efeito de publicações de notícias no mercado de
2011 ações do Brasil / Philippe Santa Maria Nizer ; orientador, Júlio César Nievola.
– 2011.
53 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2011
Bibliografia: f. 45-49

1. Ações (Finanças) - Processamento de dados. 2. Bolsa de valores. 3. Mercados. 4. Econometria. 5. Informática. I. Nievola, Júlio César. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. III. Título.

CDD 20. ed. – 004

Aos meus pais, Gilberto e Sonia, e à minha esposa, Paula.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ADBM25	*Average Document BM25*
BM25	*Best Match 25*
DF	*Document Frequency*
HME	Hipótese do Mercado Eficiente
IBOVESPA	Índice Bovespa
NB	*Naïve Bayes*
SVM	*Support Vector Machine*
TF	*Term Frequency*
DF	*Documento Frequency*
KDD	*Knowledge Discovery in Database*
KDT	*Knowledge Discovery in Text*

Resumo

A Hipótese do Mercado Eficiente diz que o valor de um ativo financeiro é dado por todas as informações disponíveis sobre ele num determinado momento. Porém, não há como um único analista financeiro estar ciente de todas as notícias referentes a um conjunto de ações no exato momento em que elas são publicadas. Assim, um sistema computacional, que aplicasse técnicas de mineração de textos para a análise do conteúdo de notícias publicadas em tempo real em conjunto com técnicas de econometria para previsão de volatilidade dos valores negociados de ativos financeiros, poderia ajudar analistas e simples investidores a selecionar quais são as notícias que causariam maior impacto no comportamento dos valores das ações em interesse. Com o auxílio de modelos de previsão de volatilidade e da capacidade crescente de processamento dos computadores, é possível descobrir se uma determinada notícia pode causar um impacto considerável nos preços de uma ação que está sendo negociada. Este trabalho tem como objetivo criar um método para se fazer a análise do conteúdo de notícias em português sobre empresas que possuem ações negociadas na bolsa de valores e tentar prever qual será o efeito destas notícias no comportamento do mercado de ações brasileiro.

Palavras-Chaves: mineração de textos, previsão de valores, mercado de ações, efeito de notícias.

Abstract

The Efficient Market Hypothesis states that the value of an asset is given by all information available in the present moment. However, there is no possibility that a single financial analyst be aware of all published news which refer to a collection of stocks in the moment that they are published. Thus, a computer system that applies text mining techniques to analyze the content of real time news, simultaneously with econometric techniques for predicting the volatility of financial assets may helps analysts and simple investors to choose which news cause the higher impact on stock market behavior. With the assistance of volatility forecast models and the growing computers processors capacity, it is possible to find out if certain news may cause a considerable impact on prices of a negotiated stock. This work has the goal of creating a method for analyzing Portuguese written news's content about companies that have their stocks negotiated in a stock market and trying to predict what kind of effect these news will cause in the Brazilian stock market behavior.

Keywords: text mining, volatility forecast, stock market, news effect.

Capítulo 1

Introdução

Nos últimos anos viu-se uma grande mudança na forma de se obter informações. Hoje, os computadores e as redes formadas por eles são a maneira mais eficiente e utilizada de se trocar informações. A Internet tornou as informações disponíveis a um número muito grande de pessoas em todo o mundo, além de possibilitar a transmissão delas de um ponto a outro do planeta em questão de segundos.

Por conseqüência disso, analistas do mercado financeiro vêem os computadores como grandes aliados na obtenção de informações que os auxiliem a atribuir valores justos às grandes quantidades de ativos com que trabalham. A informática permite que eles tenham acesso às notícias e relatórios econômicos no momento em que são publicados, para que então definam o melhor valor de compra ou de venda de determinado ativo.

Além do mais, os analistas utilizam os computadores na aplicação de modelos que tentam prever o comportamento do mercado. Com o auxílio destes modelos e da capacidade crescente de processamento dos computadores, é possível calcular com certa confiabilidade quais os riscos de se negociar um ativo dado seu comportamento histórico.

Este capítulo traz a motivação para a realização deste trabalho, além de apresentar quais são os objetivos, e uma descrição de como este documento está organizado.

1.1 Motivação

O trabalho de analistas financeiros consiste em sugerir aos seus clientes a compra de ações (ou qualquer outro tipo de ativo financeiro) que julgam abaixo do valor real e a venda daquelas ações que estariam acima do valor real. Portanto, determinar qual o valor real de um ativo é uma das principais tarefas do analista.

Com base na Hipótese do Mercado Eficiente (FAMA, 1970), o valor de um ativo é dado pelas informações disponíveis no momento. Porém, não há como um único analista financeiro ler todas as notícias referentes a um conjunto de ações no momento em que são publicadas e deste modo estar um passo a frente do comportamento do mercado. Assim, um sistema computacional que aplique técnicas de mineração de textos pode ajudar a selecionar quais são as notícias que causariam maior impacto no comportamento dos valores das ações através da análise do conteúdo de notícias.

Como mencionado anteriormente, modelos que tentam prever o comportamento do mercado são utilizados vastamente por analistas financeiros para analisar o risco de se negociar determinadas ações. Porém, poucos métodos estudados hoje tentam fundir estes dois aspectos importantes de previsão do mercado: a análise do conteúdo de notícias e os modelos econométricos existentes.

1.2 Objetivos

O objetivo do trabalho descrito neste documento é criar um método de classificação automática de notícias importantes que citem empresas as quais possuem ações negociadas na bolsa de valores. As notícias deverão ser classificadas entre “interessantes” e “não interessantes”, sendo que devem estar nesta classe as notícias que não trazem informações importantes e naquela as que trazem informações que podem ter um impacto significativo nos valores das ações negociadas.

Com esse método, será possível criar um sistema que filtre notícias irrelevantes, e mostre ao usuário apenas aquelas que tragam alguma informação importante, para que ele possa tomar a atitude conveniente. Isto é importante por que hoje o número de notícias publicadas diariamente na Internet é crescente, além de que o conteúdo criado através de *blogs* e outros meios adicionam-se ao número do conteúdo criado pelos grandes portais, e isso torna cada vez mais difícil para apenas um analista financeiro (ou até mesmo para um simples investidor) manter-se atualizado com as últimas notícias publicadas sobre uma determinada empresa.

Além disso, será necessário a utilização de uma base de dados contendo um considerável número de notícias que contenham a data e hora se sua publicação, e também o

variação do preço de ações ao decorrer do tempo. É objetivo também do trabalho a criação dessa base de dados.

Assim, espera-se com este trabalho facilitar a forma com que o usuário acompanha notícias sobre empresas descartando – ou apenas armazenando-as em uma pasta de notícias diferencias, por exemplo – as notícias que não precisam ser lidas imediatamente.

1.3 Hipótese

A hipótese que levou o desenvolvimento do presente trabalho é a seguinte: é possível criar um método que identifique automaticamente notícias importantes sobre ações de uma determinada empresa utilizando um histórico de notícias e o histórico da variação do preço das ações que contenham o período da publicação das notícias.

1.4 Contribuições Científicas e Tecnológicas

A principal contribuição desse trabalho é a adaptação de um método proposto para a classificação automática de notícias para a língua portuguesa, para que desse modo, notícias publicadas no Brasil sobre ações negociadas na Bolsa de Valores de São Paulo possam ser analisadas e classificadas. Até onde se sabe, este trabalho realizou pela primeira vez a tentativa de classificar automaticamente notícias importantes publicadas em português e relacionadas com ações negociadas na bolsa de valores. Com a realização deste, foram desenvolvidos códigos em linguagens de programação que poderão ser utilizados mais tarde em outros sistemas que visem a classificação automática de textos, por exemplo, um sistema disponível na Internet que selecione para um determinado usuário qual as notícias mais importantes para serem lidas. Além disso, foram coletadas notícias que se referem às principais empresas com ações negociadas na Bolsa de Valores de São Paulo (Bovespa) de diversos portais publicadas durante todo o ano de 2009. Todas estas notícias foram armazenadas em formato XML e poderão ser utilizadas por outros pesquisadores.

1.5 Organização do Documento

Este documento está dividido em “Temas e Trabalhos Relacionados”, onde estão expostos, primeiro, uma introdução nos assuntos abordados neste trabalho e nas técnicas de

aprendizagem e de econometria que foram utilizadas, e em seguida a apresentação de alguns estudos que realizaram de alguma forma o que foi proposto neste trabalho; “Método”, que apresenta qual foi o processo, no qual este trabalho se realizou, através da articulação das técnicas de mineração de textos e dos modelos econométricos; “Resultados Obtidos”, que traz tabelas com os resultados obtidos com a aplicação do método proposto através de tabelas, gráficos e comentários sobre estes, e “Conclusão e Trabalhos Futuros”, onde está apresentado uma síntese sobre aquilo que se fez durante o trabalho, e que se pode tirar dos resultados obtidos.

Capítulo 2

Temas e Trabalhos Relacionados

Neste capítulo serão discutidos alguns conceitos teóricos de temas sobre os quais foi preciso ter conhecimento para o desenvolvimento do método proposto. Também serão apresentadas técnicas e abordagens desenvolvidas até então, cujo objetivo é similar ao deste trabalho.

Assim, serão apresentados temas como a previsão de tendências do mercado de ações através de análise de séries temporais e mineração de textos. Para um melhor entendimento, o capítulo foi dividido em: “Mercado de Ações”, “Modelos de Previsão para Séries Temporais Financeiras”, “Descoberta do Conhecimento em Textos” e “Trabalhos Relacionados”.

2.1 Mercado de Ações

Ações são “títulos de renda variável, emitidas por sociedades anônimas, que representam a menor fração do capital da empresa emitente” (NORONHA, 2004). São amplamente negociadas por possuírem grande liquidez, ou seja, são facilmente compradas e vendidas. As empresas que vendem suas ações no mercado estão procurando uma forma barata de obter capital e, desta forma, aplicar estes novos recursos em melhorias que possibilitem o crescimento da empresa e o aumento dos lucros. Os compradores das ações participam destes lucros através dos dividendos que os acionistas têm direito. O valor de determinadas ações varia de acordo com as leis de mercado, como a lei da oferta e da procura.

Existem vários perfis de pessoas que atuam no mercado de ações: investidores de longo prazo, investidores de médio prazo, especuladores e leigos. A forma de cada um comprar e vender ações interfere no seu preço. Existem também pessoas que atuam no

mercado tentando prever o comportamento das ações, e desta maneira, identificar o momento certo de comprá-las ou vendê-las.

Prever o comportamento do mercado de ações é um grande desafio. Até recentemente, o meio acadêmico acreditava que o mercado financeiro funcionava de maneira essencialmente aleatória. Porém, pessoas que trabalham nas bolsas de valores sempre disseram ter motivos para acreditar que existem maneiras de prever o comportamento dos preços das ações.

A primeira teoria que se propunha identificar um padrão em que os preços de ações estão sujeitos foi baseada nas análises financeiras feitas por Charles Henry Dow no final do século XIX. Esta teoria, que mais tarde ficou conhecida como Teoria Dow (NORONHA, 2004), identificava três grandes movimentos (um de curto, outro de médio e o último de longo prazo) nos preços das ações, cada um deles com freqüência e tendência diferente.

No mercado financeiro, há basicamente dois tipos de análises possíveis para tentar prever o comportamento das ações. A primeira refere-se as análises feitas apenas no histórico e nos gráficos de preços das ações (a teoria Dow é um exemplo desta análise) e ela chama-se **análise técnica**. Os que analisam os dados econômicos e financeiros das empresas, juntamente com dados macro-econômicos nacionais e internacionais, fazem a **análise fundamentalista**. Porém, no meio acadêmico, surgiram teorias como o *Random Walk* (JOHNSON, 1988) que diz que toda forma de tentar prever valores futuros através da análise do histórico de preços é ineficiente.

Com o surgimento de novas técnicas econométricas e de modelos financeiros heterocedásticos[1], como o ARCH (ENGLE, 1982), a previsão do comportamento do mercado financeiro voltou a ser tema corrente das pesquisas científicas, e hoje as pessoas que trabalham com ações e outros ativos financeiros utilizam vastamente estas técnicas.

Neste capítulo serão discutidas as principais teorias existentes sobre o funcionamento aleatório ou não do mercado financeiro, a forma com que as informações sobre as bolsas de valores estão disponíveis e suas descrições.

[1] Os modelos heterocedásticos lidam com a diferença de variância existente em diferentes sub-conjuntos pertencentes a um conjunto de variáveis aleatórias. Para mais informações sobre esses modelos consultar Engle (1982).

2.1.1 Hipótese do Mercado Eficiente

De acordo com a Hipótese do Mercado Eficiente (HME), o valor de um ativo em um determinado instante de tempo reflete diretamente todas as informações disponíveis sobre este ativo no momento (FAMA, 1970). Assim, novas informações seriam responsáveis por alterar o valor de um ativo, e a competição do mercado ajustaria o preço para o seu valor real (JOHNSON, 1988). A hipótese sugere que o mercado é um sistema de equilíbrio, com o preço do ativo sempre tendendo ao seu valor real determinado pelas informações. Assim, a melhor previsão do valor de um ativo para o dia seguinte é o seu valor atual.

A partir da HME, surgem teorias que dizem que o mercado tem um comportamento essencialmente aleatório e que qualquer forma de previsão é ineficaz. Uma dessas teorias é o *Random Walk*, que diz que os preços dos ativos têm duas fontes de alterações: as provocadas pela ação imediata de compradores e vendedores, e as novas informações. A primeira é impossível de prever, pois as pessoas que compram ou vendem seus ativos o fazem por razões diversas e sem relação com a atitude de outras pessoas. A segunda é essencialmente aleatória: novas informações são incorporadas ao mercado constantemente e de maneira imprevisível, pois nunca é possível saber quando estas informações serão disponibilizadas, nem qual será o conteúdo ou importância de cada nova informação (JOHNSON, 1988).

Porém, estudos mostram que a Hipótese do Mercado Eficiente não retrata inteiramente a realidade (FAMA, 1991). Além do mais, modelos econométricos desenvolvidos a partir da década de 80 e a crescente capacidade de processamento dos computadores mostram que é possível prever com alguma confiabilidade o comportamento do mercado. Como exemplo disto, podemos citar os modelos da família ARCH (ENGLE, 1982). Mesmo assim, é evidente a importância das informações, que segundo a HME, traduzem-se em variação do valor dos ativos.

2.1.2 Informações Numéricas

As informações numéricas geradas pelo funcionamento das bolsas de valores são basicamente: o valor negociado, o volume de negócio e o *tick*. Através destas séries é possível calcular outras que trazem informações utilizadas para a análise do comportamento das ações. Cada uma destas informações são séries temporais, pois para cada instante de tempo t um valor diferente está registrado.

a) Preço: O preço de uma ação em determinado instante de tempo é o valor referente à última negociação realizada antes deste instante. Por exemplo, uma série de valores diários indica o valor da última negociação executada em cada dia da série. É possível encontrar séries que contenham três valores de preço para cada instante de tempo. São os valores máximos e mínimos de negociações que a ação atingiu entre o instante de tempo anterior e o atual e o valor da última negociação.

b) Retorno: Indica qual foi a variação do preço de uma ação com relação ao instante anterior. Geralmente é medido através de uma variação porcentual e indica qual seria o lucro ou prejuízo de um investimento feito no instante anterior. Este índice também pode ser medido como sendo variação logarítmica do preço com relação ao seu instante anterior.

c) Volume: O volume indica o número de ações que foram negociadas dentro de uma janela de tempo determinada entre o instante anterior e o instante atual de medição. Porém, ao se verificar um número elevado no volume em determinado tempo, não se pode concluir que há necessariamente uma agitação no mercado. Ocorre que é muito comum um único grande investidor negociar um grande número de ações.

d) Tick: O *tick* é a quantidade de negociações (o ato de compra/venda de um número x de ações) que foram concluídas dentro de uma janela de tempo determinada entre o instante anterior e o instante atual de medição. É o índice mais indicado para se determinar se há ou não uma agitação no mercado, pois não há diferenças entre a negociação realizada entre pequenos e grandes investidores.

e) Volatilidade: Este índice indica como a atividade do mercado varia em relação ao tempo e é, geralmente, mais útil ao investidor do que simplesmente o preço de um ativo. Porém, para se medir a volatilidade é preciso determinar uma janela de tempo compatível com a forma do investimento que se deseja realizar. Para investidores que atuam no mercado *intraday*, ou seja, que compram e vendem ações para obter lucro dentro de um único dia, a volatilidade que se deseja medir não pode levar em conta o comportamento do mercado

durante meses. Da mesma forma que um período curto de mais pode conter muito ruído (ROLL, 1984).

$$v_t = \sqrt{\frac{1}{n}\sum_{j=0}^{n-1}\left(R_{t-j}\right)^2} \quad \textbf{(2.1)}$$

A equação (2.1) descreve como a volatilidade pode ser calculada, na qual R é o retorno obtido no tempo *t* e *n* é o número de elementos de retorno que serão utilizados para o cálculo da volatilidade. Neste trabalho, onde o interesse é obter a volatilidade a mais imediata possível, será utilizado sempre $n = 1$. Portanto, a volatilidade em *t* torna o módulo do retorno em *t-1*.

2.1.3 Informações Textuais

São informações não estruturadas escritas em forma de notícias ou de análises financeiras. Têm o objetivo de informar leitores de jornais ou revistas especializadas sobre acontecimentos em geral, ou também de sugerir uma determinada ação. O objetivo destas informações é transmitir conhecimento de uma pessoa a outra através de uma linguagem corrente.

As informações referentes ao mercado financeiro podem ser divididas entre: notícias macroeconômicas, notícias gerais e recomendações de analistas. A seguir a explicação referente a cada uma delas.

2.1.3.1 Notícias Macroeconômicas

São notícias sobre economia que trazem informações sobre decisões de bancos centrais e governos. Influenciam diretamente o funcionamento das bolsas de valores, mas como geralmente ocorrem de forma agendada, o seu conteúdo pode ser antecipado pelo mercado. É possível observar um aumento na volatilidade do mercado nos dias que precedem uma notícia macroeconômica (BOMFIM, 2000).

Foi verificado que as publicações de notícias macroeconômicas não afetam o mercado por si mesmas, mas que o conteúdo das notícias é o causador das mudanças (KIM et. al., 2004).

Notícias macroeconômicas de um país podem afetar o mercado de outro (NIKKINEN et. Al, 2004). Kim (1998) observou que notícias macroeconômicas sobre os Estados Unidos ou sobre a Austrália podem afetar a volatilidade da taxa de câmbio entre o Dólar norte americano e o Dólar australiano.

2.1.3.2 Notícias Gerais

São as demais notícias disponibilizadas ao público que não são macroeconômicas. Podem conter qualquer tipo de conteúdo, desde previsão do tempo até informações sobre a compra de uma empresa por outra. As notícias gerais podem impactar diretamente na cotação de determinado ativo, ou podem ser completamente irrelevantes.

A volatilidade da série de retornos de ativos financeiros cresce no momento em que notícias comuns ligadas a estes ativos são publicadas (MELVIN e YIN, 2000).

Uma pesquisa demonstrou a relação entre notícias sobre a previsão do tempo e a variação do preço futuro de uma *commodity* (ROOL, 1984).

2.1.3.3 Recomendações de Analistas

São publicações que tem o objetivo de sugerir ao leitor uma determinada ação. Por exemplo, pode ser recomendado ao leitor que compre ações de uma determinada empresa pois através de um estudo apresentado no texto é demonstrado que o valor intrínseco da ação é superior ao valor de compra, e por isto é vantajoso realizar a compra. Pode conter recomendações para venda, ou para cautela na negociação de determinados papéis.

Geralmente são escritos por profissionais do mercado financeiro, e não são disponibilizados para o público em geral, mas somente para os clientes de empresas de corretagem ou *sites* especializados na Internet.

2.1.4 Notícias Boas ou Ruins?

É tarefa difícil classificar automaticamente notícias cujo conteúdo pode ser considerado bom ou ruim pelos analistas financeiros. De um modo simplificado, esperaríamos que se fosse publicada uma notícia positiva com relação a uma determinada empresa, o mercado, então, reagiria também de forma positiva, fazendo com que o valor da ação desta

empresa subisse. Porém, não é isto que se observa. É possível que o mercado antecipe o conteúdo de uma notícia positiva através de pessoas com fontes privilegiadas. Se a notícia publicada não é tão boa quanto aquela esperada (mas mesmo assim, positiva no conteúdo), o valor das ações poderiam cair.

2.2 Modelos de Previsão para Séries Temporais Financeiras

Existem atualmente diversas técnicas que visam prever valores de séries temporais. É comum encontrar na literatura modelos matemáticos que tentam prever valores dada uma série de valores históricos.

Porém, quando se trata de séries financeiras, os modelos que tentam prever valores começaram a ser vastamente utilizados somente depois da década de 1980, quando Engle desenvolveu o modelo ARCH. Este foi o primeiro modelo que utiliza formalmente as mudanças que ocorrem na variância das séries financeiras (BERA e HIGGINS, 1993).

Mais recentemente, trabalhos utilizando Redes Neurais e outras técnicas de reconhecimento de padrões também se mostraram capazes de prever o comportamento de séries de tempo financeiras, algumas vezes mais precisas que modelos puramente matemáticos (ROH, 2007).

O objetivo deste capítulo é descrever os modelos mais utilizados que tentam prever tendências do mercado financeiro.

2.2.1 *Moving Average* (MA)

Este é um modelo simples que tenta prever valores futuros através de uma seqüência de médias de subconjuntos da série temporal em análise (BOLLERSLEV e DOMOWITZ, 1993). Na definição da equação (2.2), x_t é a série de previsão, β_i são os pesos para cada deslocamento de tempo, e μ_t é um seqüência com médias calculadas da série original. O modelo utiliza as últimas n médias, e quanto menor é seu valor, mais rápido os valores passados são esquecidos pelo modelo.

$$x_t = \beta_0 + \mu_t + \sum_{i=1}^{n} \beta_i \mu_{t-i} \qquad \textbf{(2.2)}$$

Existem algumas variações do modelo MA como o *Assymetric Moving Average* (asMA), que divide os pesos β_i e as séries μ_t em dois tipos: um com pesos e médias para variações negativas e outro para variações positivas. Isto tenta melhorar a previsão dos valores porque se sabe que o mercado reage de forma diferente para variações positivas ou negativas (TSAY, 2005).

Outra variação é o *Exponentially Weighted Moving Average* (EWMA) proposto por Roberts (1959), que por sua vez introduz pesos na forma de $\beta = e^{-\alpha}$. Desta forma, os pesos dão mais importância para períodos mais recentes e diminui a importância dos períodos mais afastados de forma exponencial negativa, de acordo com a constante α.

2.2.2 Auto-regressivo (AR)

O modelo auto-regressivo faz o uso de recursão para tentar melhorar a previsão dos valores futuros. Assim como os modelos MA, utiliza uma seqüência que identifica as variações na série, dada pela equação (2.3), onde α_0 e α_i são constantes encontradas na maximização de uma função de verossimilhança do modelo com a série temporal efetiva.

$$x_t = \alpha_0 + \mu_t + \sum_{i=1}^{n} \alpha_i x_{t-i} \qquad \textbf{(2.3)}$$

2.2.3 Heterocedasticidade Condicional Auto-regressiva (ARCH)

Considera-se que uma série de tempo financeiro possui uma variância v. Ao dividir esta série em séries menores de determinada janela de tempo, verifica-se que a variância de cada uma dessas janelas possui uma diferença em relação à v. Verificou-se que em séries financeiras de retorno esta diferença apresenta mudanças suaves ao longo do tempo (TSAY, 2005). Os modelos heterocedásticos levam em consideração estas mudanças e esta característica permitiu que estes modelos fossem vastamente utilizados na análise de mercado financeiro.

Seja P_t o preço de uma ação no tempo t. O retorno da ação no tempo t em comparação com o período $t-1$ é dado por $r_t = \ln(P_t) - \ln(P_{t-1})$. Deste modo, podemos

definir que a média condicional (m_t) e a variância condicional (h_t) da variável aleatória r_t dado as informações disponíveis no tempo $t-1$ são:

$$m_t = E[r_t \mid \Im_{t-1}] \tag{2.4}$$

$$h_t = E[(r_t - m_t)^2 \mid \Im_{t-1}] \tag{2.5}$$

Isto implica que a série de retorno R_t é gerada através de $R_t = m_t + \sqrt{h_t}\varepsilon_t$, onde $E[\varepsilon_t \mid \Im_{t-1}] = 0$ e $V[\varepsilon_t \mid \Im_{t-1}] = 1$.

O modelo ARCH(P), proposto por Engle (1982) para modelar a volatilidade da inflação do Reino Unido, modela a volatilidade usando a definição da equação (2.6), onde α_t e α_0 são constantes de pesos, que devem ser calculadas utilizando um método de máxima verossimilhança que minimiza o erro da função através da alteração das constantes, e h_t é a volatilidade prevista no tempo *t*.

$$h_t = \alpha_0 + \sum_{i=1}^{P} \alpha_i R_{t-i}^2 \tag{2.6}$$

Este modelo assume que os retornos positivos e negativos têm o mesmo efeito sobre a volatilidade uma vez que se baseia no quadrado dos retornos anteriores.

2.2.4 ARCH Generalizado (GARCH)

Este modelo foi proposto por Bollerslev (1986) e é uma generalização do ARCH que permite variâncias condicionais e não condicionais. Em testes foi observado que o ARCH precisa de um longo período antes que a variância condicional seja alterada dentro do modelo. O GARCH permite uma mudança mais rápida desta variância (BOLLERSLEV, 1986).

O modelo GARCH(P,Q) é definido pela equação (2.7) onde *P* e *Q* são os atrasos da auto-regressão para retorno e volatilidade respectivamente; α_t, α_0 e β_j são constantes de pesos, e h_t é a volatilidade prevista no tempo *t*.

$$h_t = \alpha_0 + \sum_{i=1}^{P} \alpha_i R_{t-i}^2 + \sum_{j=1}^{Q} \beta_j h_{t-j} \quad \textbf{(2.7)}$$

2.2.5 ARCH Heterogêneo (HARCH)

Proposto por Müller et. al (1997) e é definido pela equação (2.8), onde α_t, α_0 e β_j são constantes de pesos. Este modelo possibilita que a volatilidade de longo prazo tenha um peso maior que aquela de curto prazo, tornando possível levar em consideração vários horizontes de tempo.

$$h_t = \alpha_0 + \sum_{i=1}^{P} \alpha_i \left(\sum_{j=1}^{i} \beta_j \sqrt{h_{t-j}} \times R_{t-j} \right)^2 \quad \textbf{(2.8)}$$

2.2.6 *Threshold* GARCH (TARCH)

O modelo TARCH é um modelo GARCH assimétrico. Isto quer dizer que o modelo prevê comportamento diferente para variações positivas e negativas. Ele tentar melhorar a previsão dos valores, já que o mercado reage de forma diferente para variações positivas e negativas (TSAY, 2005). Este modelo foi proposto por Glosten et. al (1993) e Zakaian (1994) e é definido pela equação (2.9), onde δ_{t-k} é uma variável de indicação que recebe o valor um se o residual no tempo $t-k$ é negativo, e, caso contrário, recebe zero; α_0, α_t, β_t e γ_t são constantes de pesos, e h_t é a volatilidade prevista no tempo *t*.

$$h_t = \alpha_0 + \sum_{i=1}^{P} \alpha_i (R_{t-i} - \mu)^2 + \sum_{j=1}^{Q} \beta_j h_{t-j} + \sum_{k=1}^{R} \delta_{t-k} \gamma_k (R_{t-k} - \mu)^2 \quad \textbf{(2.9)}$$

2.3 Descoberta do Conhecimento em Textos

O termo descoberta de conhecimento em textos é utilizado para designar todo um conjunto de técnicas que ajudam pessoas a transformar dados de texto não-estruturados em informações e conhecimento úteis (FELDMAN, 1995). Assim como no caso da descoberta do conhecimento em banco de dados (KDD), este conhecimento útil pode ser considerado como a extração não-trivial de informações implícitas, desconhecidas e potencialmente úteis (FRAWLEY, 1991).

Ao contrário de banco de dados tradicionais, onde as informações armazenadas estão organizadas de forma estruturada, os textos de notícias, relatórios, manuais, etc. são exemplos de dados não estruturados (HEARST, 2003). Por causa disso, as tradicionais técnicas de KDD não podem ser aplicadas diretamente aos textos (FELDMAN, 1995).

Após marcar os textos com informações estruturadas nas etapas de pré-processamento, é possível aplicar as técnicas tradicionais de KDD nos textos não estruturados. Estas informações estruturadas geralmente trazem as palavras e frases mais freqüentes do texto (HEARST, 2003).

Da mesma forma que mineração de dados é uma etapa do processo de descoberta do KDD, mineração de textos (TM) é uma etapa do processo de KDT. O processo de KDT é composto por vários passos, que vão desde a coleta de texto até a visualização dos dados, como se pode observar na Figura 1.

2.3.1 Classificação de Documentos

Segundo Han e Kamber (2006), “classificação é o processo de encontrar um modelo ou uma função que descreve ou distingue classes de dados ou conceitos com o propósito de usá-lo para prever a classe de objetos dos quais não se conhece o rótulo”. O modelo é criado usando como base um “conjunto de treinamento” cujos dados possuem um rótulo.

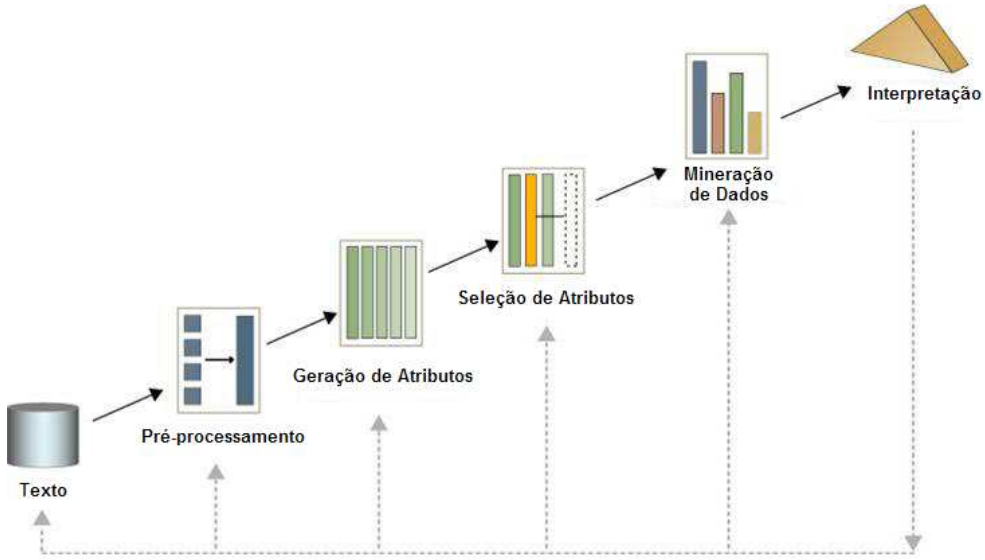

Figura 1: Processo KDT

Fonte: Even-Zohar (2002)

Existem várias técnicas de aprendizagem de máquina que são usadas com o objetivo de classificação. Uma delas é a classificação baseada em **regras**, onde os valores dos dados são testados na forma SE-ENTÃO. Outra técnica é a **árvore de decisão**, que utiliza uma estrutura de fluxo em forma de árvore onde cada nó representa um teste que deve ser feito no valor de algum atributo da instância em teste, cada tronco representa um resultado possível para o teste e cada folha representa a classe resultante da classificação (HAN e KAMBER, 2006). Porém, as técnicas mais utilizadas para classificação de documentos são o Naïve Bayes e o SVM.

2.3.2 Representação dos Documentos

Os algoritmos de classificação como o Naïve Bayes e o SVM não trabalham diretamente com um documento textual; estes algoritmos não “lêem” o texto por uma seqüência de palavras, que formam frases com sujeito, verbo e complementos. A forma comum de se trabalhar com um texto em KDT é reduzí-lo a um conjunto de dados que consiga representá-lo da melhor forma possível. Estes dados podem ser representados, por exemplo, por uma matriz que indica a quantidade de vezes que uma determinada palavra aparece num documento, como mostrado na Tabela 1. Este método de representação de um documento chama-se *Term Frequency (TF)*.

Tabela 1: Exemplo de uma matriz com valores t*erm frequency*.

	D1	***D2***	***D2***	***D4***	***D5***
palavra1	0	1	0	1	0
palavra2	0	0	2	0	0
palavra3	1	0	0	3	0
...					

Porém, o TF pode não ser suficiente para indicar a importância de uma palavra no documento. Por exemplo, se existem palavras que são muito comuns, então sua presença num determinado documento pode não ser muito importante. Mas se a palavra é rara, ela deveria ter um peso maior que as palavras mais usadas. Por isso existe também um método chamado de *Term Frequency Inverse Document Frequency* (TF-IDF), no qual busca-se medir a importância da palavra atribuindo um peso maior àquelas que são mais raras. Com este método, é calculado um valor para uma determinada palavra dentro de um documento com a equação (2.10), onde N é o número total de documentos existentes dentro do conjunto de treinamento, d_j é o *term frequency*, ou quantas vezes a palavra aparece no documento, e df_j é o *document frequency* (DF), ou o número de documentos dentro do conjunto de treinamento em que a palavra aparece.

$$TFIDF = d_j \times \log_{10}\left(\frac{N}{df_j}\right) \quad \textbf{(2.10)}$$

O TF-IDF é bastante utilizado pela comunidade nas aplicações de mineração de textos, por isso foi o método escolhido para o desenvolvimento deste projeto.

2.3.3 Seleção dos Atributos

Outro problema em mineração de textos é a seleção das palavras que representarão os documentos. Dependendo da natureza dos documentos utilizados, o número de palavras torna a tarefa de treinamento bastante custosa computacionalmente. Muitas vezes, a seleção de um

conjunto restrito de palavras pode melhorar, inclusive, o resultado alcançado pelos classificadores.

Decidiu-se neste trabalho pela criação de um conjunto restrito de palavras, as quais serão os atributos representativos de um documento (porém, serão realizados também testes sem a seleção de atributos, ou seja, utilizando todas as palavras encontradas no conjuntos dos documentos). Para isso, existem alguns métodos de seleção de palavras cujo objetivo é tentar selecionar as palavras que melhor conseguem representar os documentos de acordo com um objetivo específico. Como o objetivo deste projeto é classificar documentos, então a seleção das palavras precisa levar em consideração a capacidade que uma determinada palavra tem em separar uma classe da outra. Além disso, é preciso escolher previamente qual será o número de termos (ou palavras) que serão utilizados para a representação do documento.

Existe alguns métodos que tentam medir a eficiência com que uma determinada palavra consegue separar uma classe de outra. A seleção de atributos consiste em criar uma lista de n palavras das quais o método utilizado indique serem as mais representativas. Um dos índices para seleção de palavras mais utilizado é o *Information Gain*. Além deste, é utilizado neste projeto o ADBM25, proposto por Robertson (2008).

Information Gain

O *Information Gain* utiliza-se de um cálculo que mede qual a eficiência que uma determinada palavra tem em representar uma classe através da presença dela em cada um das diferentes classes. Por exemplo, uma palavra que aparece somente nos documentos da classe X terá uma pontuação maior que aquelas que aparecem em sua maioria na classe X, mas também na classe Y. Porém, se uma palavra aparece igualmente em todos as classes, esta terá a pontuação mínima.

A equação (2.12) mostra como é calculado o *Information Gain* para classificadores binários (que contenham apenas duas classes) do termo *j*, onde *N* é o número de documento dentro do conjunto de treinamento, df_j é o *document frequency* (DF), ou o número de documentos dentro do conjunto de treinamento em que a palavra aparece, r_j é a quantidade de documento da classe em interesse em que a palavra aparece e R é a quantidade total de documentos da classe em interesse. A equação E(n,m) é a função de entropia.

$$E(n,m) = -\left(\frac{n}{m}\log_2\left(\frac{n}{m}\right) + \left(1-\frac{n}{m}\right)\log_2\left(1-\frac{n}{m}\right)\right) \mid n \le m \quad \textbf{(2.11)}$$

$$Gain_j = E(R,N) - \frac{df_j}{N} \times E(r_j, df_j) - \frac{N-df_j}{N} \times E(df_j - r_j, df_j) \quad \textbf{(2.12)}$$

ADBM25

Este método, proposto por Robertson (2008), é uma adaptação do método BM25 (ROBERTSON e SPÄRCK JONES, 2006) e leva em consideração o tamanho dos documentos em que determinado termo aparece e também o tamanho médio dos documentos. Com este método, um termo que aparece *n* vezes em documentos pequenos tem menos importância que um termo que parece as mesmas *n* vezes em documento grandes (com maior número de palavras). Este método foi aplicado com sucesso nos trabalhos de classificação de notícia.

A equação (2.13) mostra como é calculado a eficiência de separação do termo *j*, onde *N* é o número de documento dentro do conjunto de treinamento, d_j é o *term frequency*, ou quantas vezes a palavra aparece do documento, df_j é o *document frequency* (DF), ou o número de documentos dentro do conjunto de treinamento em que a palavra aparece e , r_j é a quantidade de documento da classe em interesse em que a palavra aparece, R é a quantidade total de documentos da classe em interesse, $dl_{(i)}$ é o tamanho do documento *i* e *avdl* é o tamanho médio dos documento do conjunto de treinamento. Neste projeto, as constantes utilizadas foram $k_1 = 1$ e $b = 0{,}5$, por serem os valores utilizados por Robertson (2008).

$$ADBM25_j = \frac{1}{N}\sum_{i=1}^{N} \frac{(k_1+1)\times d_j}{k_1 \times \left((1-b) + b \times \frac{dl_{(i)}}{avdl}\right) + d_j} \times \log\left(\frac{(r_j+0{,}5)\times(N-df_j-R+r_j+0{,}5)}{(df_j-r_j+0{,}5)\times(R-r_j+0{,}5)}\right) \quad \textbf{(2.13)}$$

Após selecionar os termos que possuem as *n* melhores pontuações (tanto utilizando o *Information Gain* quanto o ADBM25), foi então criado uma matriz para cada um dos documentos selecionados com os valores de cada um dos atributos calculados com o TF-IDF.

2.4 Trabalhos Relacionados

Para realizar a classificação de notícias de acordo com um critério específico é possível utilizar técnicas de aprendizagem de máquina. Porém, o maior problema prático relacionado à aprendizagem de máquina diz respeito ao treinamento de um sistema de classificação. Essas técnicas necessitam de um grande conjunto de dados previamente classificados. Koppel e Shtrimberg (2006) resolveram este problema utilizando a variação dos dados numéricos das bolsas de valores para classificar uma notícia em boa ou ruim. Desta forma é possível treinar um sistema de classificação para reconhecer notícias boas ou más. Neste trabalho observou-se que as notícias são consideradas boas como padrão e a presença de determinadas palavras as tornam más.

Fung (2003) pesquisou uma abordagem com Mineração de Textos e múltiplas séries temporais. Neste trabalho foi desenvolvido um mecanismo para encontrar co-relações entre o valor de diferentes ações. Os resultados obtidos mostram que a análise com as múltiplas séries temporais são melhores daqueles que usam apenas uma série temporal.

Com relação a pesquisas com a movimentação financeira *intraday*, ou seja, aquelas que acontecem durante um único dia com séries temporais de pequenos intervalos de tempo (na casa dos minutos), não foi possível encontrar um número significativo de publicações. As mais importantes foram as realizadas por Marc-André Mittermayer e Calum Robertson.

Mittermayer criou um sistema que classifica notícias entre boas, ruins e neutras. Para marcar as notícias, criou uma forma de analisar os valores da ação nos 60 minutos após a publicação da notícia utilizando limiares. Após o treinamento de um sistema de classificação, obteve uma taxa de acerto de 58% (MITTERMAYER, 2004).

Robertson (2006) concluiu que apenas uma fração das notícias é responsável pelas reações mais significativas do mercado. Em um trabalho posterior (ROBERTSON, 2007a) ele investigou os efeitos que as notícias causam no mercado *intraday* analisando séries temporais como retorno, valor, volatilidade e o modelo GARCH (*Generalised Autoregressive Conditional Heteroskedasticity*). No trabalho seguinte (ROBERTSON, 2007b), notícias foram marcadas com a etiqueta de “interessante” quando a diferença do erro da previsão do modelo de volatilidade (GARCH) com o encontrado realmente é maior que um limiar. Treinando um sistema de classificação, obteve precisão de mais de 80% e concluiu que o mercado reage

rapidamente às notícias (cerca de 5 minutos depois da publicação). Robertson (2008) também propôs maneiras de melhorar o modelo GARCH utilizando a análise de notícias. A Tabela 2 mostra os resultados obtidos por Robertson na classificação de documentos. TP (*True Positive*) são os documentos que causaram alguma alteração no mercado classificados corretamente, FP (*False Positive*) são os documentos que não causaram nenhuma alteração no mercado e foram classificados como se tivessem causado, FN (*False Negativo*) são documentos que causaram alteração no mercado mas foram classificados como se não tivessem causados, e TN (*True Negative*) são aqueles que não causaram mudança nenhuma e foram classificados corretamente.

Tabela 2: Melhores resultados obtidos por Robertson (2008).

País	***Class.***	***TR***	***Termos***	***Total***	***TP***	***FP***	***FN***	***TN***	***Sens.***	***Espec.***	***Prec.***
EUA	SVM	GAIN	1000	133019	973	27769	1859	102418	34,36%	78,67%	77,73%
ING	C4.5	GAIN	100	81522	348	6559	1442	73173	19,44%	91,77%	90,19%
AUS	SVM	ADBM25	5000	33098	331	4856	515	27396	39,13%	84,94%	83,77%
EUA	SVM	ADBM25	100	133019	674	25270	921	106154	42,26%	80,77%	80,31%
ING	C4.5	GAIN	100	81522	308	8687	895	71632	25,60%	89,18%	88,25%
AUS	SVM	ADBM25	2000	33098	215	4538	365	27980	37,07%	86,04%	85,19%

A Tabela 2 mostra os resultados obtidos por Robertson (2008), na qual os dados estão separados por País — Estados Unidos (EUA), Inglaterra (ING) e Austrália (AUS) —, por tipo de classificador (SVM ou C4.5), por método de seleção de termos (GAIN ou ADBM25). A tabela trás o número de termos selecionados para a representação dos documentos e o total de documentos utilizados nos testes. Os resultados estão separados em TP (verdadeiro positivo), FP (falso positivo), FN (falso negativo) e TN (verdadeiro negativo). Os índices medidos são: sensibilidade (Sens.), que é a proporção de documentos que causam alteração no mercado classificados corretamente, a especificidade (Espec.) que é a proporção de documentos que não causam alteração no mercado classificados corretamente, e a precisão (Prec.) que é a proporção total de documentos classificados corretamente.

Capítulo 3

Método

Neste capítulo será descrito o método utilizado para a realização do presente trabalho, como os dados utilizados foram obtidos, como foi feito o pré-processamento dos dados textuais, quais foram as técnicas utilizadas para a rotulação automática dos dados, quais os métodos utilizados para a seleção de atributos e para o treinamento de um classificador automático.

3.1 Base de Dados

Neste capítulo se descreve como foram coletadas as informações necessárias para a realização do trabalho. Os dados necessários foram as notícias textuais e as séries temporais de valores negociados das ações da bolsa de valores.

3.1.1 Índice Bovespa

O Índice Bovespa (IBOVESPA) é um portfólio de ações elaborado pela Bolsa de Valores de São Paulo. Pertencem a este portfólio as ações que são negociadas na Bovespa e são consideradas mais as importantes no período da elaboração do índice. Este índice não é estático, pois há alterações periódicas tanto na lista de ações que pertencem ao índice, como no peso que cada ação tem dentro do portfólio.

Por serem consideradas as principais empresas que têm suas ações negociadas na Bovespa, é possível que as empresas que fazem parte do Ibovespa sejam mais citadas na imprensa. Por este motivo que se decidiu utilizar este índice como referência para a escolha das empresas cujas ações serão objeto de análise.

As notícias coletadas foram aquelas publicadas durante o ano de 2009 e que citassem as seguintes empresas que possuem suas ações negociadas na Bovespa e que fazem parte do Ibovespa: ALL, Ambev, Bradesco, Brasil Telecom, Braskem, Celesc, Cemig, Cesp, Comgas, Copel, Cosan, Cyrela, Eletrobras, Eletropaulo, Embraer, Gerdau, Gol, Itaú, Klabin, Light, Natura, Net, Petrobrás, Sabesp, Siderúrgica Nacional, Souza Cruz, Tam, Telesp, TIM, Usiminas, Vale e Vivo. O motivo da escolha dessas empresas é o fato delas participarem no índice criado pela Bovespa (o Ibovespa) que reúne as ações com maior volume de negociações, ou seja, são as principais ações.

3.1.2 Informações Textuais

As notícias coletadas foram as que citaram pelo menos uma vez o nome da empresa, e que possuem o horário de sua publicação. Foram utilizadas notícias em português de veículos online, ou seja, os portais da Internet que publicam notícias continuamente.

Para a busca das notícias de determinado período foi utilizado o Google News Archive, um sistema do Google que possibilita a busca de notícias de um determinado período de tempo. Foi desenvolvido um programa que, através da lista de links resultantes de uma busca realizada no Google News Archive, baixava automaticamente as páginas HTML que continham as notícias. De cada uma das páginas HTML, foi separado o texto, a data de publicação, a manchete e o nome do veículo de publicação e armazenado dentro de um arquivo XML. Os portais de notícias utilizados para a aquisição dos dados textuais foram: Folha Online, O Estadão, O Globo e Valor Online.

Na Tabela 3 estão os números de notícias publicadas durante o ano de 2009 e que foram utilizadas como base de dados para a execução deste trabalho.

3.1.3 Informações Numéricas

Foram coletadas as séries de preços das ações citadas acima, com intervalo de cinco minutos entre cada cotação, e que compreendem os meses de setembro, outubro, novembro e dezembro de 2009. A fonte destes dados foi o sistema de acompanhamento do mercado financeiro Bloomberg Terminal.

Os arquivos contendo as cotações foram armazenados no formato CSV (*Comma-separeted Values*) e possuem, além do valor da ação para cada período de cinco minutos, um campo que informa a data e horário da cotação.

Tabela 3: Base de dados de notícias por empresa.

Empresa	Notícias	Empresa	Notícias
ALL	62	Gol	640
Ambev	212	Itaú	1.525
Bradesco	1.472	Klabin	184
Brasil Telecom	484	Light	53
Braskem	215	Natura	379
Celesc	79	Net	25
Cemig	184	Petrobrás	5.414
Cesp	99	Sabesp	446
Comgas	71	Siderúrgica Nacional	159
Copel	97	Souza Cruz	82
Cosan	212	Tam	900
Cyrela	257	Telesp	98
Eletrobras	535	Tim	1.172
Eletropaulo	348	Usiminas	562
Embraer	661	Vale	147
Gerdau	611	Vivo	156
Total	17.541		

3.1.4 Pré-processamento

As notícias armazenadas em modo HTML precisaram passar por uma etapa de pré-processamento que removeu os dados indesejados dos arquivos como cabeçalhos, rodapés, propagandas, *tags* HTML de imagens, *links*, tabelas e formatação de texto. Para esta etapa, foi desenvolvido um programa que ao ler as arquivos XML que continham as notícias, removia as *tags* HTML e também substituía os códigos para caracteres especiais pelos caracteres Unicode específicos.

Após este processo, de cada uma das notícias armazenada em HTML resultaram apenas um texto puro, sem formatação, imagens ou efeitos, que então poderá ser utilizado nos processos de remoção de *stop words* e de *stemmer* das palavras.

Stop Words são aquelas palavras consideradas irrelevantes para um processo de mineração de textos por conta de sua repetição e também por não caracterizar algo

objetivamente, como são os artigos, preposições etc. A remoção dessas palavras visa diminuir o universo de dados que necessitam de processamento, e que, por sua vez, as palavras que contenham algum significado relevante possam ser mais facilmente encontradas. Para a realização desse trabalho, foi utilizado uma lista de *stop words* da língua portuguesa disponível livremente no site Linguateca (http://www.linguateca.pt/).

Em seguida, o texto das notícias passaram pelo processo de s*temmer* (ou radicalização) que tem por objetivo agrupar todas as palavras que possuem o mesmo radical numa representação única. Assim, depois deste processo, toda variação de palavras que possuem o mesma origem, e portanto, o mesmo radical, passam então a ser consideradas como a mesma. Além disso, neste processo, toda a alteração que uma palavra sofre em sua terminação (como número, gênero para os substantivos e a conjugação nos diversos tempo e pessoas para os verbos) também são eliminados (VIEIRA e VIRGIL, 2007).

Existem poucos algoritmos para a radicalização de palavras da língua portuguesa. Um deles é uma modificação para o português do algoritmo de Porter (1980), e outro, também baseado neste, conhecido como algoritmo de Orengo (2004). Este algoritmo baseia-se na ação dos seguintes passos:

1. Remoção dos sufixos;
2. Remoção dos sufixos verbais, se o primeiro passo não realizou nenhuma alteração;
3. Remoção do sufixo i, se precedido de c;
4. Remoção dos sufixos residuais os, a, i, o, á, í, ó;
5. Remoção dos sufixos e, é, ê e tratamento da cedilha.

Para este trabalho, foi utilizado uma implementação em Java chamada PTStemmer (OLIVEIRA, 2010) do algoritmo de Orengo.

3.2 Rotulação das Notícias

Para o treinamento de um método de classificação é necessária a existência de um conjunto de dados previamente classificados. Porém, os textos obtidos dos portais de notícias

não possuem nenhuma rotulação prévia. É preciso, então, que estes textos sejam rotulados seguindo algum critério.

Por isto, baseado no processo utilizado por Robertson (2008), decidiu-se por rotular os textos entre duas classes: “interessantes” e “não interessantes”. Assim como Robertson, para se decidir se uma notícia é importante ou não, foi utilizado a análise da dados contidos nas séries temporais de cotações das ações da empresa citada na notícia.

Se uma notícia causa algum impacto no preço da ação, espera-se que nos momentos seguintes à publicação da notícia encontre-se uma movimentação “anormal”. Essa movimentação não esperada pode ser observada num aumento da volatilidade do preço da ação. Ou seja, nos momentos seguintes, espera-se que a variação entre valorização e desvalorização do preço da ação se intensifique de alguma maneira. A volatilidade pode ser calculada como sendo a variância do retorno (valorização ou desvalorização) de uma ação. Pode ser calculada pela equação (3.1), onde R_t é o retorno verificado no tempo *t*. Para este trabalho, será utilizado $n = 1$.

$$\upsilon = \sqrt{\frac{1}{n}\sum_{j=0}^{n-1}\left(R_{t-j}\right)^2} \qquad \textbf{(3.1)}$$

Existem modelos econométricos que tentam prever a volatilidade de algum ativo financeiro. Algum desses métodos, como o GARCH (BOLLERSLEV, 1986), baseia-se na idéia de que um período que sucede outro de alta volatilidade tende a ter também uma alta volatilidade.

Para se decidir se uma notícia é importante ou não, primeiro é necessário que se saiba se a notícia causou alguma modificação no comportamento do mercado ou não. Foi, então, aplicado o modelo GARCH nas séries temporais de preço e depois, comparando-se a volatilidade prevista pelo modelo com a que se observou efetivamente, foi possível saber se o mercado estava operando num regime de normalidade ou não. Caso se verificasse este último caso, segundo a Hipótese do Mercado Eficiente (FAMA, 1970), pode-se considerar que o mercado está absorvendo novas informações. Portanto, se existe uma notícia publicada momento antes dessa alteração do mercado, é possível que as informações novas estejam presentes nesta notícia, e por isso, esta notícia seria marcada como “interessante”. As outras notícias são marcadas por padrão como “não interessante”.

A Figura 2 mostra o processo de rotulação da seguinte forma: através das séries de preços históricos de cada uma das ações são obtidas duas outras séries: a séries de volatilidade real e a séries de volatilidade prevista por um modelo; no passo seguinte é obtida uma terceira série que representa o erro entre a série de volatilidade real e a série de volatilidade prevista; com a lista de notícias publicadas, obtém-se da série de erro qual foi o valor que ocorreu nos momentos próximos da publicação da notícia; se o erro da volatilidade passar de um determinado limiar, rotula-se a notícia como “interessante”.

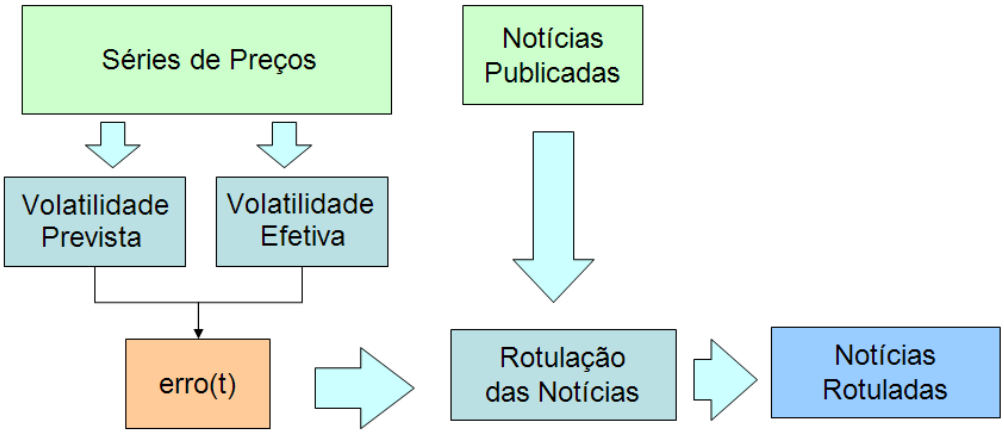

Figura 2: Rotulação das Notícias

A Figura 3 mostra graficamente os dados que representam o valor de uma ação durante um período de tempo e o horário de publicação de uma notícia. Uma forma de avaliar a importância de uma notícia é utilizar um modelo de volatilidade e calcular o erro entre o valor previsto e a volatilidade efetiva. Como os modelos não são capazes de prever uma alteração causada externamente, o erro representará o grau do impacto que um evento não previsto causou na volatilidade.

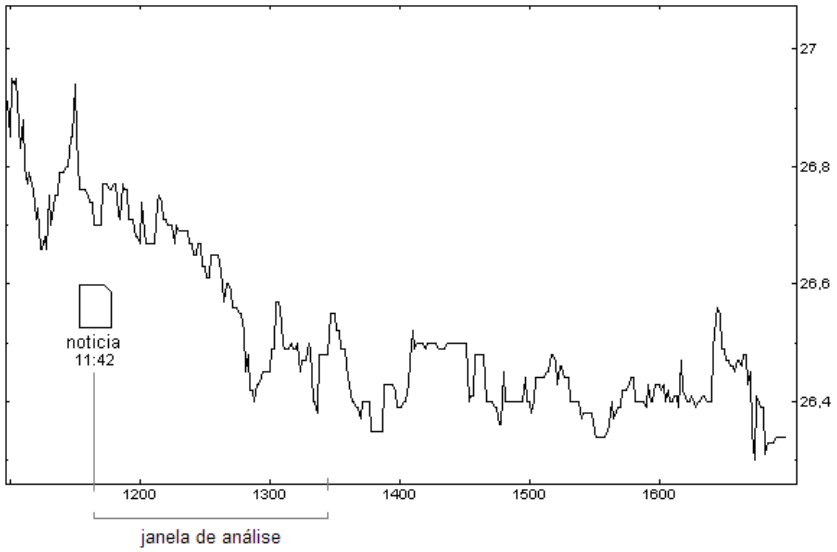

Figura 3: Publicação de Notícia e Série de Preços

Robertson (2007a) demonstrou que o erro entre a previsão da volatilidade de uma ação usando o modelo GARCH e seu valor real tem uma alta relação com a publicação de notícias sobre aquela ação. Em outro trabalho, Robertson (2007b) categorizou notícias pelo erro do modelo GARCH em dada janela de tempo. Se o erro ultrapassasse o desvio padrão com relação à média, a notícia era marcada como “interessante”. A média foi calculada no início de cada dia pelo erro dos últimos 20 dias de negociação.

Este trabalho também utiliza o erro entre a previsão de volatilidade e o valor. Para cada mês que se deseja classificar as notícias, foi utilizado um modelo GARCH(3,3) que tem seus parâmetros calculados com os dados referentes ao mês anterior. É calculado, então, um valor de erro médio e sua variância com os valores de volatilidade do mês anterior e aqueles valores obtidos com o uso do modelo. Para se rotular uma notícia como “interessante”, é preciso que se identifique um período de tempo no qual a diferença entre a volatilidade efetiva e a prevista seja maior que a média do erro histórico somado a sua variância. Assim, se uma notícia está dentro de uma janela de tempo $\Delta\tau$ do instante em que foi identificada essa “anormalidade”, ela é rotulada como “interessante”.

Foi utilizada uma série de valores para $\Delta\tau$, e para cada um deles foi medido a eficácia do resultado da classificação automática de notícias. Os valores em minutos utilizados para $\Delta\tau$ foram: 5, 10, 15, 20 e 30. Estes valores foram escolhidos levando-se em conta que

Robertson (2007a) mostrou que a reação do mercado a publicação de notícias reflete rapidamente na volatilidade.

3.3 Recuperação de Informações

Após a rotulação das notícias, foi necessário selecionar os dados de cada um dos textos para o treinamento de um classificador automático. Primeiro é preciso selecionar quais são as palavras de cada uma das notícias que serão utilizadas como atributos representativos da notícia. É preciso ter muito cuidado para a seleção dessas palavras, pois palavras importantes não podem ser descartadas, assim como palavras que não ajudarão na classificação precisam ser ignoradas.

Foram testados diversos métodos de seleção de atributos. Foram eles: *information gain* e ADBM25, que estão descritos no capítulo 2.3. Foi utilizado uma função TF-IDF para representar os valores de cada um dos atributos das notícias. Para cada variação dos métodos de seleção dos atributos foi medida a eficácia dos classificadores automáticos.

3.4 Treinamento do Classificador

Os métodos de classificação utilizados na pesquisa foram o Naïve Bayes e o SVM por serem os mais adequados para classificação de documentos textuais (DUDA, 2002). Foi utilizado o sistema Weka, que contém os algoritmos de classificação mencionados, e que também faz testes com o classificador treinado (WEKA).

3.5 Medição de Resultados

Para medir os resultados dos classificadores serão utilizados os seguintes critérios: TP (*True Positive*) indica quantos documentos causaram alguma alteração no mercado (“interessantes”) foram classificados corretamente; FP (*False Positive*) indica a quantidade de documentos que não causaram nenhuma alteração no mercado (“não interessante”) e foram classificados como se tivessem causado; FN (*False Negativo*) é a quantidade de documentos que causaram alteração no mercado mas foram classificados como se não tivessem causados; TN (*True Negative*) é a quantidade de documentos que não causaram mudança nenhuma e foram classificados corretamente; e N é a quantidade total de documentos. Além disso, serão

utilizados os medidores de: Sensibilidade, que é a proporção de documentos “interessantes” classificados corretamente; Especificidade, que é a proporção de documentos “não interessantes” classificados corretamente; e Precisão, que é a proporção total de documentos classificados corretamente.

$$Sensibilidade = \frac{\#TP}{\#TP + \#FN}$$

$$Especificidade = \frac{\#TF}{\#TN + \#FP}$$

$$Precisão = \frac{\#TP + \#TN}{N}$$

Capítulo 4

Resultados Obtidos

Neste capítulo, serão apresentados os resultados obtidos na execução das tarefas descritas no capítulo anterior. Também será feita uma análise dos resultados, bem como as justificativas das decisões tomadas durante a execução do projeto. Este capítulo divide-se em "Rotulação de Notícias", "Seleção dos Atributos" e "Classificador de Notícias por Empresa".

4.1 Rotulação de Notícias

Foi criado um software escrito na linguagem de programação Java para realizar o processo de rotulação das notícias descrito em 3.2. Mas antes disso, notou-se que seria necessário excluir da lista de notícias aquelas que não foram publicadas nos horários de funcionamento da bolsa. A princípio, as notícias publicadas fora do horário de funcionamento da bolsa poderiam ser acumuladas e classificadas de acordo com o os primeiros movimentos do dia seguinte de negociações. Porém, o número de notícias publicadas em um final de semana, por exemplo, seriam classificadas da mesma maneira, misturando, assim, notícias que deveriam estar em classes diferentes. Isto foi feito para que as notícias que contenham os elementos necessários para serem consideradas importantes (ou seja, que contenham aquelas palavras que refletem sua futura classificação como "interessante"), mas que não fossem publicadas nos horários que não se pode identificar seus efeitos, não influenciem negativamente no processo de treinamento dos classificadores. Deste modo, do total das notícias contidas no banco de dados, foram eliminadas neste primeiro processo aquelas que não fazem parte dos meses em análise e as que não foram publicadas em horário comercial. Na Tabela 4 estão os números obtidos de notícias por empresas depois da filtragem inicial. Depois desse primeiro processo, foi executado o processo de rotulação variando-se o valor atribuído a $\Delta\tau$ em 5, 10, 15, 20 e 30 minutos. Para cada uma das variações nota-se uma

alteração no número de notícias classificadas entre “interessante” e “não interessante”. Na Tabela 5 está a quantidade de notícias obtidas para cada classificação.

É importante observar que o número de notícias marcadas como “interessante” é bastante reduzido comparado com o número de notícias normais (“não interessante”). Com Δt=30 a proporção entre as duas classes é de 0,221 “interessante” para cada “não interessante”. Para Δt=5 esta proporção chega a 0,049.

Tabela 4: Número de notícias utilizadas para criação de um classificador.

Empresa	***Notícias***	***Empresa***	***Notícias***
ALL	6	Gol	99
Ambev	37	Itaú	201
Bradesco	216	Klabin	28
Brasil Telecom	48	Light	17
Braskem	37	Natura	59
Celesc	20	Net	4
Cemig	31	Petrobrás	761
Cesp	18	Sabesp	91
Comgas	20	Siderúrgica Nacional	29
Copel	31	Souza Cruz	9
Cosan	31	Tam	143
Cyrela	49	Telesp	35
Eletrobras	64	Tim	148
Eletropaulo	48	Usiminas	51
Embraer	108	Vale	10
Gerdau	81	Vivo	22
Total	2.552		

Tabela 5: Notícias “interessantes” e “não interessantes” por janela de tempo (Δt).

Classe	***Δt=5***	***Δt=10***	***Δt=15***	***Δt=20***	***Δt=30***
Interessante	120	211	286	355	463
Não interessante	2.432	2.341	2.266	2.197	2.089

Como mencionado, o número de notícias classificadas como “interessante” é consideravelmente inferior ao número de “não interessante”. Esse era o resultado esperado, pois o que se verifica é que não são todas as notícias que trazem informações novas, as quais são capazes de ocasionar uma movimentação mais intensa no mercado; a maioria das notícias

trazem informações redundantes ou pontuais, que não geram maiores conseqüências na volatilidade das ações.

Em testes preliminares que utilizaram os classificadores automáticos, e em cujos dados de treinamento se tentava manter a proporção entre notícias “interessantes” e “não interessantes”, observou-se uma tendência em classificar a maioria das notícias como “não interessantes”. Por causa disso, decidiu-se que seria utilizada nos conjuntos de treinamento dos classificadores uma proporção fixa entre as notícias “não interessante” e “interessantes” de 2:1, ou seja, para cada duas notícias normais (“não interessante”), foi adicionado no conjunto de treinamento uma notícia “interessante”. Em todos os testes apresentados neste trabalho, foi utilizada essa proporção.

A Tabela 6 traz exemplos de como alguns documentos foram classificados. A janela de tempo utilizada neste teste foi de 5 minutos ($\Delta t=5$) e as notícias se referem a empresa **Brasil Telecom**.

Tabela 6: Exemplo de Notícias classificadas da Brasil Telecom.

Data	***Manchete***	***Classe***
30/12/2009 12:40	Bolsa divulga novo Ibovespa para 2010; empresas de Eike são as novidades	interessante
12/11/2009 12:19	Net ultrapassa Telefônica em número de assinantes de banda larga	interessante
18/9/2009 16:31	Dólar fecha a R$ 1,80; Bovespa ascende 0,35%	interessante
26/10/2009 11:43	Bovespa começa com alta de 0,85% e dólar cai a R$ 1,706	interessante
28/12/2009 16:21	CVM multa dois diretores da Brasil Telecom em R$ 400 mil	não interessante
16/12/2009 14:19	Mais duas empresas de Eike Batista entrarão no Ibovespa	não interessante
6/12/2009 10:02	Boom pós-crise deve gerar mais gigantes	não interessante
4/12/2009 15:56	Setor de telefonia concentra mais da metade de reclamações sobre call centers	não interessante
4/12/2009 14:15	Movimento financeiro da Bovespa recua 20% em novembro	não interessante
2/12/2009 12:18	Oi lidera ranking de reclamação de consumidores no país, diz governo	não interessante
2/12/2009 10:39	Dois terços das ações da Bovespa rendem mais de 100% no ano	não interessante
1/12/2009 13:43	Empresa de logística de Eike entra no Ibovespa em janeiro	não interessante
18/11/2009 11:23	BNDES aprova financiamento de R$ 4,4 bi para a Oi	não interessante
13/11/2009 11:55	Oi pretende emitir até R$ 3 bilhões em debêntures	não interessante
6/11/2009 15:30	"Guardian" chama ex-prefeito de Curitiba Jaime Lerner de "revolucionário verde"	não interessante
23/10/2009 14:52	Bovespa tem perdas moderadas em dia cheio; dólar atinge R$ 1,71	não interessante
23/10/2009 13:34	Bovespa inverte tendência e perde 0,50%; dólar vale R$ 1,71	não interessante
23/10/2009 11:09	Bovespa avança 1% após abertura; dólar bate R$ 1,71	não interessante

4.2 Seleção dos Atributos

O conjunto dos documentos utilizados neste trabalho apresentaram um número relativamente pequeno de atributos (em todo o conjunto de documentos o número palavras após *stemmer* e excluindo-se as *stop words* não passou de 4500). Porém, para testar se um número mais restrito de palavras poderiam influenciar positivamente nos resultados, optou-se por utilizar algumas técnicas de seleção de atributos. Há diversas métodos utilizados para a seleção dos termos, dentre eles o *Information Gain* é o mais utilizado. Um outro método é o ADBM25, proposto por Robertson (2008), o qual ele faz uso na sua pesquisa sobre classificação de notícias.

Para este trabalho, foram realizados testes utilizando tanto o *Information Gain* como o ADBM25. Primeiro, foram selecionadas aleatoriamente notícias para fazerem parte do conjunto de treinamento. Para este conjunto, foram escolhidas aleatoriamente 2/3 do total de notícias “interessantes”, e das notícias “não interessante” foram escolhidas duas vezes o número total das primeiras. Em seguida foi realizada a seleção dos atributos com cada um dos dois métodos, e para cada um deles foram criados diferentes conjuntos com variação no número de termos selecionados. Foram utilizados para o número de termos os valores de 100, 200, 500, 1000, 2000 e 4000. Em seguida realizou-se os testes dos classificadores com as notícias que ficaram de fora do conjunto de treinamento (os testes foram realizados com um conjunto onde a proporção entre as notícias “interessantes” e “não interessantes” não foi modificada). Para cada variação de parâmetros, foram realizadas cinco execuções do método proposto com a respectiva medição dos resultados. Cada uma das execuções foi realizada com diferentes seleções de notícias para os conjunto de treinamento e de teste, sendo que os resultados obtidos podem ser conferidos no apêndice deste documento.

A Tabela 7 mostra alguns resultados escolhidos (os demais resultados foram desconsiderados por apresentarem sensibilidade próxima a zero). Primeiro, para cada método de seleção de atributos, a tabela mostra os resultados com melhor precisão para os dados positivos. Em seguida o melhor resultado da precisão total. O resultado na tabela mostra a média da sensibilidade (acertos das notícias “interessantes”), especificidade (acerto das notícias “não interessantes”) e a precisão do classificador das diferentes medições.

Tabela 7: Precisão dos classificadores de notícias.

	Classificador	*Δt*	*Termos*	*Sens.*	*Especif.*	*Precisão*
Gain	SVM	10	100	2,25%	97,12%	**93,96%**
	SVM	5	200	**44,50%**	60,39%	60,11%
	NB	10	100	**65,07%**	36,87%	37,80%
ADBM25	SVM	15	100	17,50%	86,36%	**83,03%**
	SVM	30	4.000	**45,16%**	68,49%	66,27%
	NB	15	100	30,21%	73,63%	**71,52%**
	NB	10	1.000	**48,73%**	54,64%	54,45%

Sensibilidade (Sens.) é a proporção de documentos "interessantes" classificados corretamente. Especificidade (Especif.) é a proporção de documentos "não interessantes" classificados corretamente. Precisão é a proporção total de documentos classificados corretamente.

Nota-se na tabela um índice de acerto relativamente baixo para as notícias "interessantes". Em apenas um dos casos, o acerto superou 50%. Nos gráficos abaixo, é possível visualizar os resultados e tentar encontrar neles alguma relação entre a variação dos parâmetros e a eficácia dos classificadores. Nos gráficos da página 36, é possível ver os resultados obtidos para os classificadores SVM, nos quais os termos das notícias foram selecionados com o método *Information Gain*. Nos gráficos seguintes (das páginas 38 e 39), repete-se a variação de classificador, mas agora com o uso do método ADBM25.

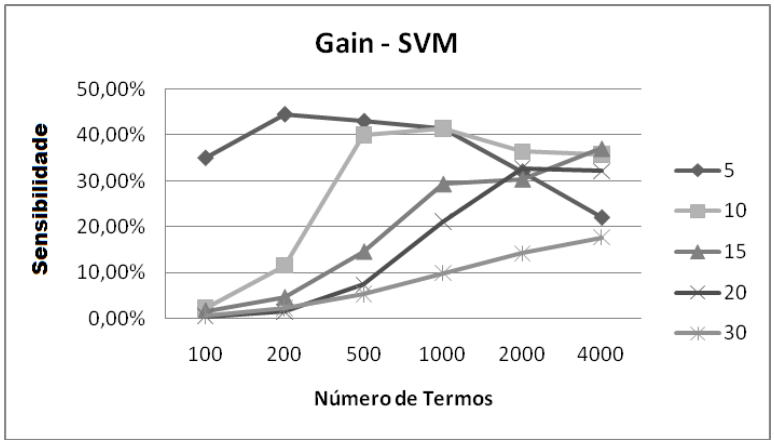

Figura 4: Sensibilidade na classificação de notícias – *Information Gain* e SVM.

Esse gráfico mostra o efeito na precisão da classificação correta de notícias "interessantes" variando-se o número de termos escolhidos para a representação de um documento e a janela de tempo Δt minutos (5, 10, 15, 20 e 30).

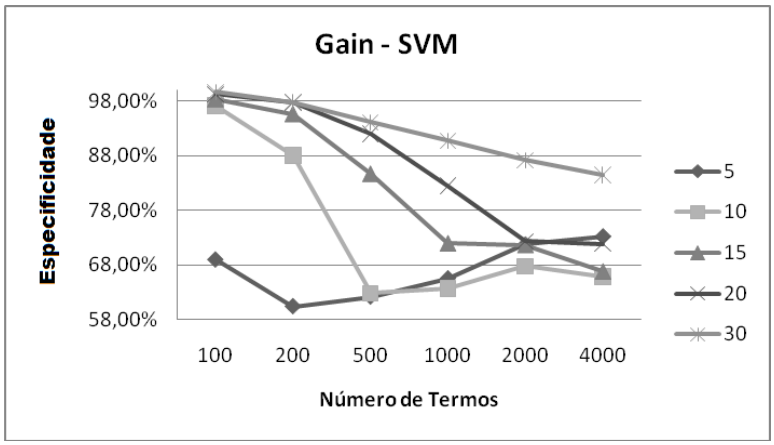

Figura 5: Especificidade na classificação de notícias – *Information Gain* e SVM.

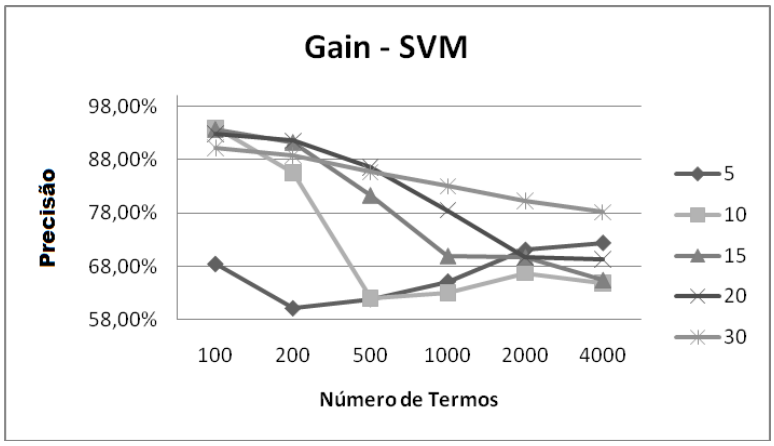

Figura 6: Precisão na classificação de notícias – *Information Gain* e SVM.

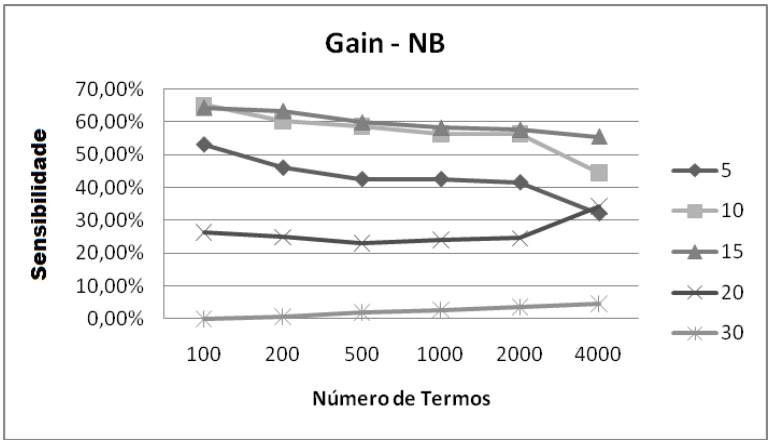

Figura 7: Sensibilidade na classificação de notícias – *Information Gain* e Naïve Bayes.

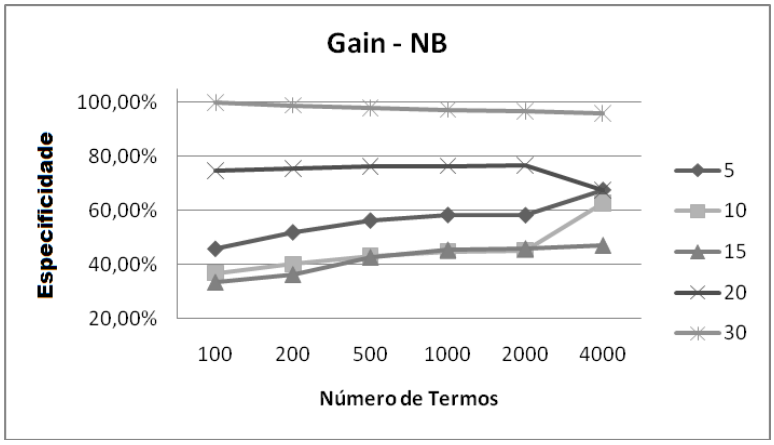

Figura 8: Especificidade na classificação de notícias – *Information Gain* e Naïve Bayes.

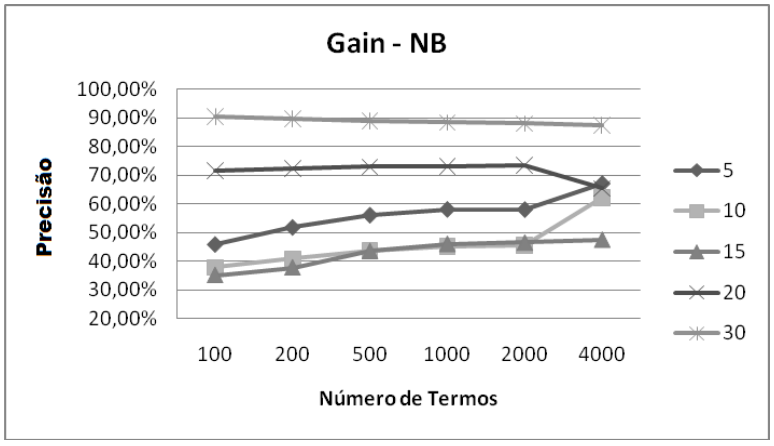

Figura 9: Precisão na classificação de notícias – *Information Gain* e Naïve Bayes.

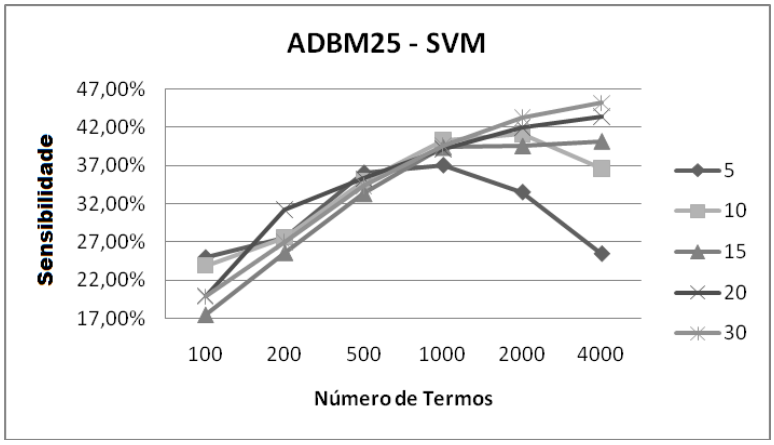

Figura 10: Sensibilidade na classificação de notícias – ADBM25 e SVM.

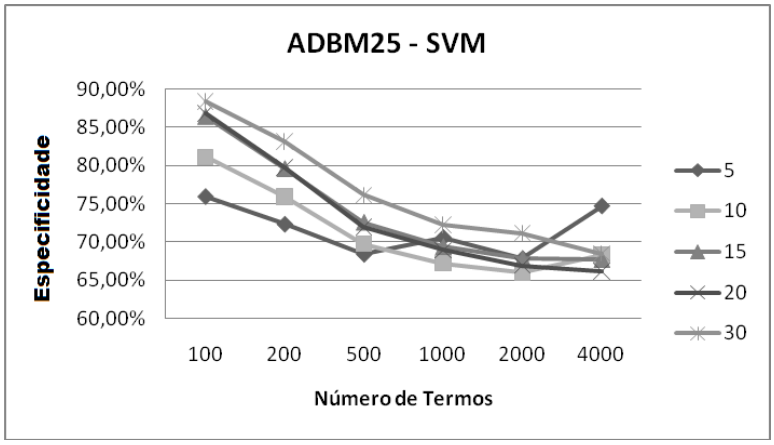

Figura 11: Especificidade na classificação de notícias – ADBM25 e SVM.

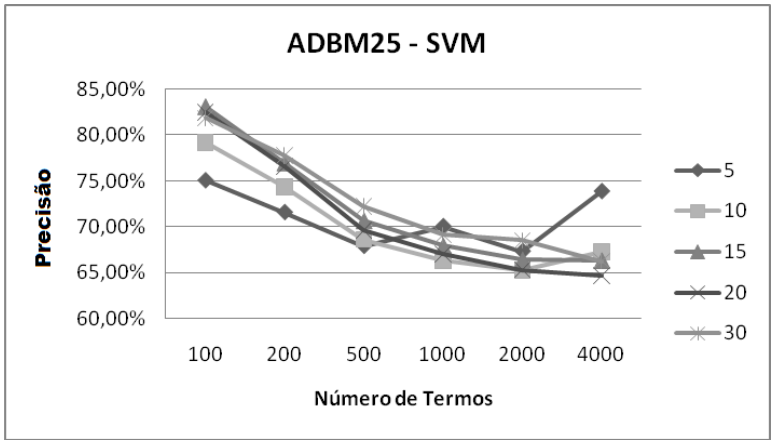

Figura 12: Precisão na classificação de notícias – ADBM25 e SVM.

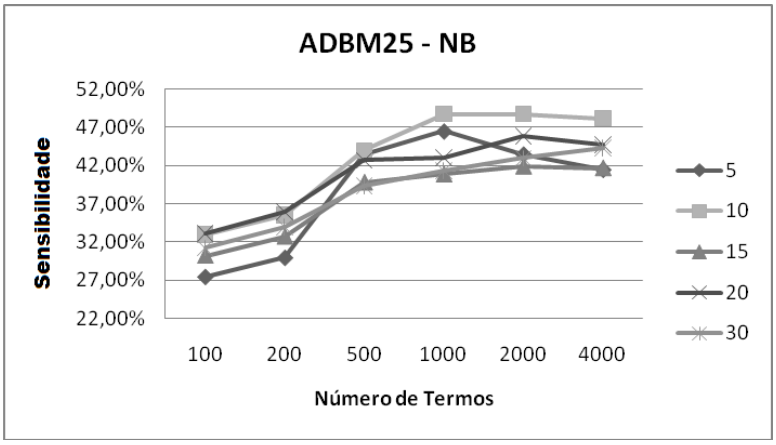

Figura 13: Sensibilidade na classificação de notícias – ADBM25 e Naïve Bayes.

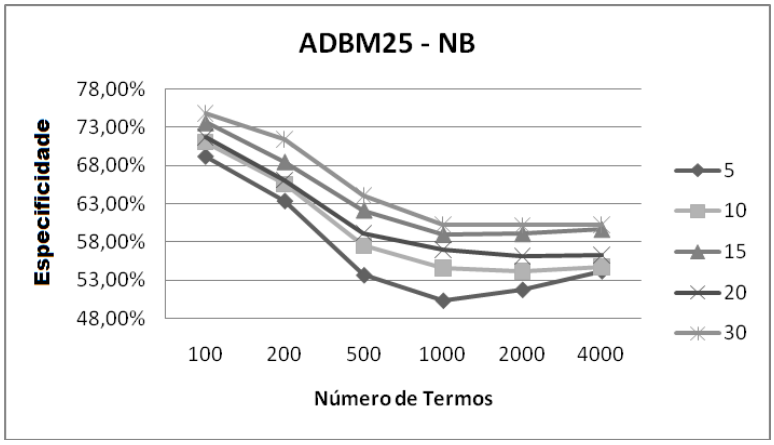

Figura 14: Especificidade na classificação de notícias – ADBM25 e Naïve Bayes.

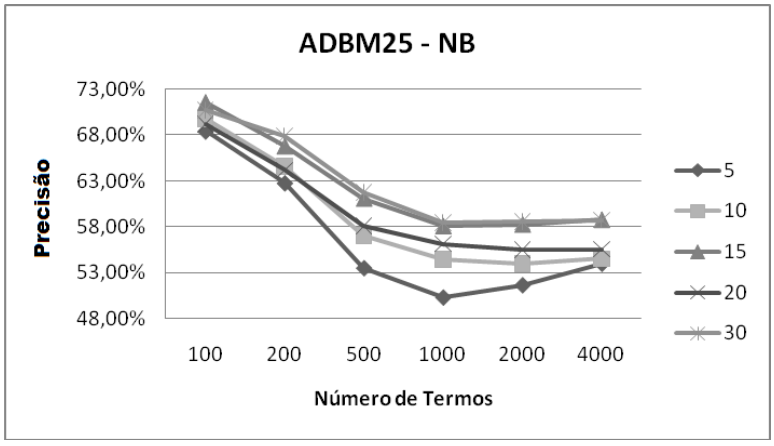

Figura 15: Precisão na classificação de notícias – ADBM25 e Naïve Bayes.

Nos gráficos é possível observar uma tendência geral, na qual a precisão para as notícias “interessantes” é pequena quando o número de termos é pequeno, e tende a aumentar quando o número de termos selecionados para o treinamento vai crescendo. A exceção desse caso é quando utiliza-se *Information Gain* para seleção de termo e o classificador Naïve Bayes. Com um número de 100 termos (o menor número testado) e janela de tempo de 15 minutos, obteve-se o melhor resultado para classificação correta de notícias “interessante”: uma média de 65,07% de acerto com um desvio padrão de 3 pontos percentuais. Apesar do melhor índice de acerto obtido para classe específica de notícias “interessante”, a precisão na classificação das outras notícias é bastante baixo: apenas 36,87%. Porém, se o objetivo do classificador de notícias for criar um filtro para um usuário final, é mais importante não perder as notícias importantes do que reduzir ao mínimo o número de notícias filtradas. No caso dos testes com o classificador Naïve Bayes com *Information Gain*, observa-se que há uma mudança menos expressiva no índice de acertos devido ao crescimento do número de termos, quando comparada com os outros classificadores e métodos de seleção de termos.

Quando nos melhores resultados para a classificação de notícias “interessantes” (*Information Gain* e Naïve Bayes – 5 e 10 termos) se encontram acertos de mais de 50%, o índice de acertos para notícias “não interessantes” é sempre menor do que isso.

Nos testes realizados com o método de seleção de termos *Information Gain*, quando se aumenta o número de termos selecionados , há uma tendência decrescente da sensibilidade (Figura 4 e Figura 7). O inverso se observa nos testes com o método ADBM25: em ambos os casos (SVM e Naïve Bayes), há uma tendência crescente na sensibilidade quando se aumenta o número de termos selecionados (Figura 10 e Figura 13). Com o método ADBM25, também se observa uma evolução mais comportada na sensibilidade quando se aumenta o número de termos (apesar de esse índice nunca ultrapassar 50%). A especificidade comporta-se sempre no sentido inverso da sensibilidade, não havendo, portanto, casos em que ambas apresentem melhoras simultaneamente.

Em todos os testes realizados, não há casos nos quais ambas as classes são classificadas com acerto maior do que 50%. Mesmo com a precisão dos classificadores sendo maior do que esse valor em quase todas as variações – em apenas alguns casos do Naïve Bayes com o *Information Gain*, a precisão do classificador não ultrapassa 50% (Tabela 11, no apêndice) –, não é possível concluir que o classificador foi bem sucedido. Isso porque existindo uma grande diferença na quantidade de notícias “interessantes” e “não

interessantes", sendo estas muito mais numerosas do que aquelas, a sensibilidade (classificação correta de "interessantes") interfere muito pouco na precisão dos classificados. Assim, a classificação correta de notícias "não interessantes" irá contribuir quase que exclusivamente para a precisão dos classificadores.

Se considerar-se o melhor caso aquele no qual a especificidade é maior do que 50% com a maior sensibilidade obtida, poder-se-ia atribuir essa qualidade aos classificadores Naïve Bayes e ADB25 com janela de tempo de 10 minutos. Na variações de quantidade de termos nos valores de 1000, 2000 obteve-se uma sensibilidade de 48,73%. A Tabela 8 apresenta os números obtidos para o classificador em questão.

Tabela 8: Resultados obtidos com classificador NB, ADBM25 e Δt = 10.

Termos	*Sens.*	*DP Sens.*	*Espec.*	*DP Espec.*	*Precisão*	*DP Prec.*
100	32,96%	0,073591	71,05%	0,071339	69,78%	0,068151
200	35,49%	0,091495	65,56%	0,080833	64,56%	0,076644
500	43,94%	0,037793	57,51%	0,068949	57,05%	0,066215
1000	48,73%	0,027456	54,64%	0,0645	54,45%	0,061938
2000	48,73%	0,025586	54,16%	0,06025	53,98%	0,057684
4000	48,17%	0,018364	54,73%	0,060238	54,51%	0,057818

Assim, destacam-se entre as variações de métodos de classificação, seleção de termos, e também dos parâmetros os classificadores:

- Naïve Bayes com *Information Gain*, 100 termos e Δt = 10 min., tendo alcançado a melhor sensibilidade (65,07%) de todos os testes realizados; ou seja, este algoritmo acerta a classificação de notícias "interessantes" 65,07% das vezes (Figura 7). Porém, o acerto do algoritmo para notícias "não-interessantes" (especificidade) é de apenas 36,87%;
- Naïve Bayes com ADBM25, 1000 termos e Δt = 10 min., tendo alcançado sensibilidade de 48,73% e especificidade de 54,64% (Tabela 8).

4.3 Classificador de Notícias por Empresa

Decidiu-se também testar qual seria a precisão dos classificadores quando eles utilizassem apenas notícias de uma determinada empresa. Por isso foram selecionadas as empresas que possuem o maior número de notícias isoladamente e então os dados foram

submetidos aos testes. Não foram feitos testes com as janelas de tempo Δt=5 pois o número de notícias “interessantes” era muitas vezes inferior a 10 (exceto para a Petrobras, que possuí o maior número das notícias coletadas por empresa).

Tabela 9: Precisão dos classificadores de notícias separadas por empresa.

Empresa		***Δt***	***Termos***	***Sensibilidade***	***Especif.***	***Precisão***
Bradesco	SVM	20	100	53,0%±8,2	61,78%±4,3	61,23%±3,8
Itaú	SVM	15	100	72,2%±16,7	63,68%±12,1	64,18%±1
Petrobrás	SVM	10	100	48,00%±13,9	60,14%±9,8	59,87%±9,3
	NB	30	500	50,50%±9,7	57,65%±5,7	54,33%±5,0
Tam	SVM	10	100	53,33%±23,3	65,49%±11,8	65,2%±11,2

Sensibilidade é a proporção de documentos “interessantes” foram classificados corretamente. Especificidade (Especif.) é a proporção de documentos “não interessantes” classificados corretamente. Precisão é a proporção total de documentos classificados corretamente.

Os resultados com classificadores Naïve Bayes, na maior parte dos casos, apresentam um desvio padrão maior do que 20 na média da sensibilidade. Para os testes apresentados na Tabela 9, utilizou-se classificadores com a seleção dos termos feito pelo método *Information Gain*; foram selecionados os dados mais estáveis, ou seja, aqueles que possuíam o menor desvio padrão da média da série dos resultados obtidos.

Capítulo 5

Conclusão e Trabalhos Futuros

Este trabalho descreve um método que tem como objetivo a identificação de notícias publicadas sobre empresas que podem causar alguma alteração no preço das ações negociadas na bolsa de valores. De acordo com um trabalho realizado para notícias publicadas em inglês no Reino Unido, Estados Unidos e Austrália (ROBERTSON, 2008), a precisão alcançada girou em torno de 80%. Porém, o dado importante para este trabalho é a sensibilidade, ou seja, o índice de acerto na classificação de notícias “interessantes” (aquelas que causam alguma alteração no mercado), pois caso seja criado um filtro para um usuário final, é importante que não se perca as notícias “interessantes”. Neste caso, o trabalho de Robertson teve o melhor resultado de 42,26% com especificação de 80,77% e precisão de 80,31%.

Os resultados obtidos neste trabalho, para notícias publicadas em português, podem se comparar com os obtidos por Robertson. O valor da precisão dos classificadores que apresentam melhor desempenho nesse índice passa de 80%. Porém, nestes casos, a sensibilidade não passa de 20%.

Para os testes realizados somente com notícias de uma única empresa, os resultados foram melhores. Isso mostra que os termos que indicam que uma notícia é “interessante” variam de acordo com o ramo de atuação da empresa, ou seja, existem termos específicos para cada uma delas. No caso da Petrobrás, os resultados não destoaram muito daqueles obtidos com o conjunto total de empresas. Isso indica que, como sendo a empresa que possui o maior número de notícias, a Petrobrás não depende de um pequeno número de termos, e, tratando-se de uma importante empresa estatal, ela é freqüentemente mencionada no noticiário político, e não apenas no financeiro.

Portanto, fica demonstrado que o método proposto é relativamente eficiente para a identificação de notícias “interessantes” e “não-interessantes” (comparando-se com os

resultados obtidos por Robertson), e que, deste modo, um filtro que aplique esta técnica poderá diminuir consideravelmente o número de notícias necessária para um investidor analisar enquanto negocia suas ações.

Também verificou-se a eficiência do método de seleção de termos ADBM25. Pelos gráficos obtidos com esse método, é possível ver uma linha mais comportada na variação dos resultados quando altera-se o número de termos.

Trabalhos futuros nesta área poderão estudar se é possível fazer a melhoria de algum modelo de volatilidade através do uso deste classificador, ou tentar aplicar algum método diferente de classificação que possa apresentar melhores resultados.

Referências

BERA, A. K.; HIGGINS, M. L.; *ARCH Models: Properties, Estimation and Testing*. Journal of Economic Surveys, Vol. 7, 1993, p 305-66.

BOLLERSLEV, T.; DOMOWITZ, I.; *Trading Patterns and Prices in the Interbank Foreign Exchange Market*. Journal of Finance, Vol. 48, 1993, p. 1421-1443.

BOLLERSLEV, T.; *Generalized Autoregressive Conditional Heteroskedasticity*. Journal of Econometrics, Vol. 31, 1986, p. 307-27.

BOMFIM, A. N.; *Pre-announcement effects, news, and volatility: monetary policy and the stock market.* Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, 2000.

DACAROGNA, M. M.; MÜLLER, U. A., PICTET, O. V.; OLSEN, R. B.; *Modelling Short-Term Volatility with GARCH and HARCH Models*. Working Paper Series, 1997, Disponível em: http://ssrn.com/abstract=36960.

DUDA, R., HART, P., STORK, D.; *Pattern Classification*. 2ed. Willey Interscience, 2002.

ENGLE, R. F. *Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation.* Econometrica: Journal of the Econometric Society, 1982, p. 987-1007.

EVEN-ZOHAR, Y. *Introduction to Text Mining, Part II. Presentation 2 of a 3-part* University of Illinois, National Center for Supercomputing Applications, Disponível em algdocs.ncsa.uiuc.edu/PR-20021116-2.ppt. Acesso em 28 de outubro de 2008.

FAMA, E. F. *Efficient Capital Markets: A Review of Theory and Empirical Work*. Papers and Proceedings of the Twenty-Eighth Annual Meeting of American Finance Association, Journal of Finance, 25(2), 1970, p. 383-417.

FAMA, E. F. *Efficient Capital Markets: II*. Journal of Finance, 46(5), 1991, p. 1575-1617.

FELDMAN, R.; DAGAN, I. *Knowledge discovery in textual databases (KDT)*. Montreal, Proc. 1st International Conference on Knowledge Discovery (KDD-95), 1995, p. 112–117.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C.J. *Knowledge Discovery in Databases: an Overview*. Knowledge Discovery in Databases, MIT Press, 1991, p. 1-27.

FUNG, G.P.C.; YU, J.X.; LAM, W. *Stock Prediction: Integrating Text Mining Approach Using Real-time News*. IEEE Int. Conference on Computational Intelligence for Financial Engineering. Hong Kong, 2003, p. 395-402.

GLOSTEN, L R; JAGANNATHAN, R.; RUNKLE, D. E.; *On the relation between the expected value and the volatility of the nominal excess returns on stocks*. Journal of Finance, Vol. 48, 1993, p. 1779–801.

HAN, J.; KAMBER, M.; *Data Mining: Concepts and Techniques*, 2 ed, Morgan Kaufmann, EUA, 2006.

HEARST, M. *What is Text Mining?*. UC Berkeley School of Information. Disponível em www.sims.berkeley.edu/~hearst/text-mining.html. Acesso em 20 de outubro de 2008.

JOHNSON, M. A. *The Random Walk and Beyond*. John Wiley & Sons Inc. 1988

KIM, S. J.; *Do Australian and the US macroeconomic news announcements affect the USD/AUD exchange rate?* Some evidence from E-GARCH estimations. Journal of Multinational Financial Management, Vol. 8, No. 2-3, 1998, p. 233-248.

KIM, S.J.; McKENZIE, M. D.; FAFF, R. W.; *Macroeconomic news announcements and the role of expectations: evidence for US bond, stock and foreign exchange markets*. Journal of Multinational Financial Management, Vol. 14, No. 3, 2004, p. 217-232.

KOPPEL, M.; SHTRIMBERG, I. *Good News or Bad News? Let the Market Decide*. The Information Retrieval Series. Vol. 20, 2006, p. 297-301.

LEE, C.M.C., MYERS, J., SWAMINATHAN, B. *What is the intrinsic value of the Dow?*, Journal of Finance, Vol. 54, 1999, p. 1693-741.

MELVIN, M.; YIN, X.; *Public information arrival, exchange rate volatility, and quote frequency*. Economic Journal, Royal Economic Society, Vol. 110(465), 2000, p. 644-661.

MITTERMAYER, M.-A. *Forecasting Intraday Stock Price Trends with Text Mining Techniques*. 37th Annual Hawaii Int. Conference on System Sciences (HICSS). Big Island, 2004, p. 64.

MÜLLER, U. A.; DACAROGNA, M. M.; DAVE, R. D.; OLSEN, R. B.; PICTET, O. V.; von WEIZSACKER, J. E.; *Volatilities of different time resolutions — analyzing the dynamics of market components*, Journal of Empirical Finance, Vol. 4, no. 2-3, 1997, p. 213-239.

NIKKINEN, J.; SAHLSTR, P.; *Scheduled domestic and US macroeconomic news and stock valuation in Europe*. Journal of multinational financial management, Vol 14, No. 3, 2004, p. 201-215.

NORONHA, M. *Curso básico de análise gráfica. Aula 1 de 2*. 2004. Disponível em http://www.timing.com.br/download/AG01.pdf. Acesso em 15 de maio de 2008.

OLIVEIRA, P.; *PTStemmer – A Stemming toolkit for the Portuguese language*. 2010. Disponível em http://code.google.com/p/ptstemmer. Acesso em 11 de fev. de 2011.

ORENGO, V. M.; HUYCK, C.; *A Stemming Algorithm for the Portuguese Language*: String Processing and Information Retrieval (SPIRE 2001) Proceeding. 8th International Symposium 13-15 Nov 2001, Chile, 2001, p. 186-193.

PORTER, M. F.; *An algorithm for suffix stripping*. Program, 14 no. 3, July 1980, p. 130-137.

ROBERTS, S. W.; *Control Chart Tests Based on Geometric Moving Averages.* Technometrics, Vol. 1, No. 3, 1959, p. 239-250.

ROBERTSON, C. S.; GEVA, S.; WOLFF, R. C. The Intraday Effect of Public Information: Empirical Evidence of Market Reaction to Asset Specific News from the US, UK, and Australia. SSRN Working Paper Series, 2007a.

ROBERTSON, C.; GEVA, S.; WOLFF, R. C. *Can the Content of Public News Be Used to Forecast Abnormal Stock Market Behaviour?*. ICDM 2007. Seventh IEEE International Conference, 2007b, p. 637-642.

ROBERTSON, C.; GEVA, S.; WOLFF, R.; *What Types of Events Provide the Strongest Evidence that the Stock Market is Affected by Company Specific News*, Fifth Australian Data Mining Conference (AusDM2006), 2006, p. 145-153

ROBERTSON, C.; *Real time financial information analysis.* Queensland University of Technology, 2008.

ROBERTSON, C.; SPÄRCK JONES, K.; Simple, Proven Approaches to Text Retrieval. University of Cambridge Computer Laboratory Technical Report no. 356, 2006.

ROH, T. H.; *Forecasting the Volatility of Stock Price Index*. Expert Systems with Applications, vol. 33, no. 4, 2007, p. 916-922.

ROLL, R.; *Orange juice and weather*. The American Economic Review, Vol. 74, 1984, p. 861-880.

SANTOS FILHO, E. L.; *Previsão dos Retornos do Índice Bovespa Usando Redes Neurais Artificiais*. PUC-PR, Dissertação de Mestrado, Engenharia de Produção e Sistemas, 2008.

TSAY, R. S.; *Analysis of Financial Time Series*, 2 ed, New Jersey: Wiley Series In Probability and Statistics, New Jersey, USA, 2005.

WEKA: *Data Mining Software in Java*. The University of Waikato. Disponível em http://www.cs.waikato.ac.nz/ml/weka/.

ZAKAIAN, J. M.; *Threshold heteroskedastic models*. Journal of Economic, Dynamics Control, Vol. 18, 1994, p. 931–55.

Apêndice

A seguir, as tabelas com a média dos resultados obtidos através dos testes realizados com os classificadores de documentos.

Tabela 10: Resultados obtidos com classificador SVM e *Information Gain*.

Janela	*Termos*	*Sens.*	*DP Sens.*	*Espec.*	*DP Espec.*	*Precisão*	*DP Prec.*
5	100	35,00%	0,212132	69,01%	0,145523	68,43%	0,139819
5	200	44,50%	0,083666	60,39%	0,029759	60,11%	0,028013
5	500	43,00%	0,069372	62,18%	0,0476	61,85%	0,04618
5	1000	41,50%	0,065192	65,54%	0,051228	65,12%	0,049513
5	2000	32,00%	0,102164	71,78%	0,070892	71,09%	0,068174
5	4000	22,00%	0,069372	73,27%	0,084649	72,39%	0,082444
10	100	2,25%	0,016059	97,12%	0,02669	93,96%	0,025578
10	200	11,55%	0,141337	88,06%	0,132924	85,52%	0,123954
10	500	40,00%	0,025586	62,80%	0,041636	62,05%	0,040224
10	1000	41,41%	0,027456	63,71%	0,051334	62,96%	0,049601
10	2000	36,34%	0,070704	67,77%	0,054525	66,73%	0,05091
10	4000	35,77%	0,055092	65,84%	0,044677	64,84%	0,044456
15	100	1,67%	0,018923	98,36%	0,019417	93,67%	0,017729
15	200	4,58%	0,026146	95,64%	0,045115	91,23%	0,042487
15	500	14,58%	0,17477	84,70%	0,142328	81,30%	0,127707
15	1000	29,38%	0,173899	71,99%	0,104748	69,93%	0,091918
15	2000	30,42%	0,12312	71,70%	0,082252	69,70%	0,073561
15	4000	37,08%	0,088572	66,84%	0,109034	65,40%	0,099719
20	100	0,50%	0,011274	99,21%	0,010919	92,84%	0,009701
20	200	1,51%	0,009205	97,73%	0,018471	91,52%	0,017377
20	500	7,39%	0,028744	91,94%	0,054516	86,49%	0,04942
20	1000	21,18%	0,210789	82,46%	0,158251	78,50%	0,134571
20	2000	32,77%	0,135109	72,30%	0,08232	69,75%	0,069607
20	4000	32,10%	0,131453	71,90%	0,122499	69,33%	0,10668
30	100	0,65%	0,007902	99,62%	0,004644	90,20%	0,003597
30	200	2,32%	0,009784	97,80%	0,027936	88,71%	0,024728
30	500	5,42%	0,029068	94,08%	0,043606	85,64%	0,037111
30	1000	9,94%	0,04196	90,70%	0,047954	83,01%	0,040795
30	2000	14,19%	0,040548	87,16%	0,049909	80,21%	0,042752
30	4000	17,68%	0,050718	84,51%	0,041768	78,14%	0,036856

Sensibilidade (Sens.) é a proporção de documentos “interessantes” classificados corretamente. Especificidade (Espec.) é a proporção de documentos “não interessantes” classificados corretamente. Precisão é a proporção total de documentos classificados corretamente. DP é o desvio padrão observado.

Tabela 11: Resultados obtidos com classificador NB e *Information Gain*.

Janela	*Termos*	*Sens.*	*DP Sens.*	*Espec.*	*DP Espec.*	*Precisão*	*DP Prec.*
5	100	53,13%	0,114337	45,86%	0,112249	45,99%	0,108337
5	200	46,00%	0,108397	51,95%	0,139331	51,85%	0,135451
5	500	42,50%	0,126244	56,30%	0,143236	56,06%	0,13874
5	1000	42,50%	0,113192	58,35%	0,142928	58,07%	0,138652
5	2000	41,50%	0,120675	58,29%	0,140759	58,00%	0,136341
5	4000	32,00%	0,020917	67,58%	0,086803	66,96%	0,085055
10	100	65,07%	0,037793	36,87%	0,060573	37,80%	0,057362
10	200	60,28%	0,047136	40,19%	0,058004	40,86%	0,054713
10	500	58,59%	0,055092	43,23%	0,061767	43,74%	0,058432
10	1000	56,34%	0,055451	44,86%	0,063199	45,24%	0,05938
10	2000	56,34%	0,055451	45,26%	0,064325	45,63%	0,060479
10	4000	44,51%	0,035352	62,67%	0,070407	62,06%	0,067163
15	100	64,38%	0,062239	33,61%	0,063697	35,10%	0,058925
15	200	63,33%	0,04963	36,47%	0,069243	37,77%	0,06446
15	500	59,79%	0,074244	42,74%	0,080738	43,56%	0,075077
15	1000	58,13%	0,082483	45,44%	0,090927	46,05%	0,084242
15	2000	57,50%	0,08149	45,93%	0,088637	46,49%	0,082295
15	4000	55,42%	0,089426	47,13%	0,082011	47,53%	0,073917
20	100	26,22%	0,353211	74,75%	0,336733	71,62%	0,292214
20	200	24,87%	0,334059	75,55%	0,310015	72,28%	0,26847
20	500	23,03%	0,264713	76,34%	0,268214	72,90%	0,233898
20	1000	24,03%	0,252423	76,52%	0,252355	73,13%	0,219893
20	2000	24,54%	0,24725	76,74%	0,23958	73,37%	0,208254
20	4000	34,29%	0,243437	67,54%	0,244186	65,39%	0,212877
30	200	0,52%	0,008412	98,95%	0,015589	89,58%	0,013377
30	500	1,81%	0,026365	98,00%	0,026838	88,85%	0,021867
30	1000	2,71%	0,028635	97,34%	0,031016	88,33%	0,025467
30	2000	3,61%	0,030123	96,86%	0,030657	87,99%	0,024942
30	4000	4,52%	0,039508	96,02%	0,037421	87,31%	0,03025

Sensibilidade (Sens.) é a proporção de documentos “interessantes” classificados corretamente. Especificidade (Espec.) é a proporção de documentos “não interessantes” classificados corretamente. Precisão é a proporção total de documentos classificados corretamente. DP é o desvio padrão observado.

Tabela 12: Resultados obtidos com classificador SVM e ADBM25.

Janela	*Termos*	*Sens.*	*DP Sens.*	*Espec.*	*DP Espec.*	*Precisão*	*DP Prec.*
5	100	25,00%	0,05863	75,90%	0,070437	75,02%	0,068671
5	200	27,50%	0,079057	72,33%	0,062847	71,56%	0,061118
5	500	36,00%	0,126984	68,45%	0,043408	67,89%	0,043114
5	1000	37,00%	0,041079	70,62%	0,075212	70,03%	0,074099
5	2000	33,50%	0,10247	67,91%	0,040402	67,32%	0,039293
5	4000	25,50%	0,054199	74,69%	0,067827	73,84%	0,066938
10	100	23,94%	0,041063	81,10%	0,041872	79,19%	0,041367
10	200	27,61%	0,027456	75,95%	0,039608	74,34%	0,038663
10	500	34,93%	0,077401	69,77%	0,028079	68,61%	0,028313
10	1000	40,28%	0,040577	67,24%	0,026006	66,34%	0,025967
10	2000	41,13%	0,068567	66,07%	0,040524	65,24%	0,040168
10	4000	36,62%	0,068277	68,34%	0,032665	67,28%	0,032743
15	100	17,50%	0,053825	86,36%	0,064726	83,03%	0,060162
15	200	25,63%	0,056385	79,55%	0,061801	76,94%	0,05749
15	500	33,33%	0,038976	72,56%	0,05844	70,66%	0,05493
15	1000	39,38%	0,054824	69,37%	0,048108	67,92%	0,043961
15	2000	39,58%	0,052602	67,82%	0,048033	66,45%	0,044007
15	4000	40,21%	0,066536	67,68%	0,029164	66,35%	0,027669
20	100	20,00%	0,055233	86,78%	0,088359	82,47%	0,081352
20	200	31,26%	0,038692	79,71%	0,071476	76,58%	0,067348
20	500	35,29%	0,048638	71,97%	0,084643	69,60%	0,077764
20	1000	39,16%	0,056246	68,95%	0,081304	67,03%	0,074431
20	2000	42,02%	0,058824	66,93%	0,07895	65,33%	0,070398
20	4000	43,36%	0,049928	66,18%	0,068981	64,71%	0,063707
30	100	19,87%	0,043614	88,35%	0,069074	81,83%	0,059156
30	200	26,97%	0,055761	83,11%	0,058423	77,76%	0,049288
30	500	34,32%	0,05328	76,17%	0,04766	72,19%	0,041605
30	1000	39,61%	0,039927	72,22%	0,051817	69,12%	0,045331
30	2000	43,23%	0,033833	71,17%	0,049364	68,51%	0,044106
30	4000	45,16%	0,050182	68,49%	0,03776	66,27%	0,034889

Sensibilidade (Sens.) é a proporção de documentos “interessantes” classificados corretamente. Especificidade (Espec.) é a proporção de documentos “não interessantes” classificados corretamente. Precisão é a proporção total de documentos classificados corretamente. DP é o desvio padrão observado.

Tabela 13: Resultados obtidos com classificador NB e ADBM25.

Janela	*Termos*	*Sens.*	*DP Sens.*	*Espec.*	*DP Espec.*	*Precisão*	*DP Prec.*
5	100	27,50%	0,068465	69,15%	0,06721	68,43%	0,066089
5	200	30,00%	0,05863	63,35%	0,073866	62,78%	0,072878
5	500	43,50%	0,118057	53,71%	0,076769	53,53%	0,074272
5	1000	46,50%	0,136473	50,40%	0,069112	50,33%	0,06631
5	2000	43,50%	0,124499	51,81%	0,075899	51,67%	0,072798
5	4000	41,50%	0,120675	54,23%	0,072449	54,01%	0,069451
10	100	32,96%	0,073591	71,05%	0,071339	69,78%	0,068151
10	200	35,49%	0,091495	65,56%	0,080833	64,56%	0,076644
10	500	43,94%	0,037793	57,51%	0,068949	57,05%	0,066215
10	1000	48,73%	0,027456	54,64%	0,0645	54,45%	0,061938
10	2000	48,73%	0,025586	54,16%	0,06025	53,98%	0,057684
10	4000	48,17%	0,018364	54,73%	0,060238	54,51%	0,057818
15	100	30,21%	0,044804	73,63%	0,080266	71,52%	0,077055
15	200	32,71%	0,040745	68,49%	0,075593	66,76%	0,071939
15	500	39,79%	0,037122	62,16%	0,072426	61,08%	0,069005
15	1000	40,83%	0,06886	58,94%	0,07129	58,06%	0,067077
15	2000	41,88%	0,077392	59,08%	0,071254	58,24%	0,066604
15	4000	41,67%	0,053623	59,63%	0,062621	58,76%	0,059328
20	100	33,11%	0,079411	71,69%	0,072315	69,20%	0,070433
20	200	35,92%	0,054622	66,12%	0,092624	64,17%	0,090042
20	500	42,69%	0,053611	59,18%	0,080079	58,11%	0,076641
20	1000	43,03%	0,073835	57,04%	0,077353	56,14%	0,073159
20	2000	45,88%	0,079411	56,20%	0,077587	55,53%	0,07304
20	4000	44,71%	0,073595	56,30%	0,077907	55,55%	0,073092
30	100	31,23%	0,053319	74,83%	0,043722	70,68%	0,042799
30	200	33,94%	0,051734	71,45%	0,061381	67,87%	0,059279
30	500	39,35%	0,060003	64,09%	0,054126	61,73%	0,052264
30	1000	41,29%	0,061035	60,26%	0,054834	58,45%	0,051012
30	2000	42,97%	0,065572	60,20%	0,056076	58,56%	0,051935
30	4000	44,26%	0,0775	60,26%	0,056773	58,73%	0,053446

Sensibilidade (Sens.) é a proporção de documentos “interessantes” classificados corretamente. Especificidade (Espec.) é a proporção de documentos “não interessantes” classificados corretamente. Precisão é a proporção total de documentos classificados corretamente. DP é o desvio padrão observado.