UNIVERSIDADE DE SÃO PAULO
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE
DEPARTAMENTO DE ADMINISTRAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM ADMINISTRAÇÃO

RECUPERAÇÃO DE INFORMAÇÃO DE COMUNICADOS À IMPRENSA

Daniel Bittencourt Silva

Orientador: Prof. Dr. Cesar Alexandre de Souza

SÃO PAULO
2011

Prof. Dr. João Grandino Rodas
Reitor da Universidade de São Paulo

Prof. Dr. Reinaldo Guerreiro
Diretor da Faculdade de Economia, Administração e Contabilidade

Prof. Dr. Adalberto Américo Fischmann
Chefe do Departamento de Administração

Prof. Dr. Lindolfo Galvão de Albuquerque
Coordenador de Programa de Pós-Graduação em Administração

DANIEL BITTENCOURT SILVA

RECUPERAÇÃO DE INFORMAÇÃO DE COMUNICADOS À IMPRENSA

Dissertação apresentada ao Departamento de Administração da Faculdade de Economia, Administração e Contabilidade da Universidade de São Paulo como requisito parcial para obtenção do título de Mestre em Administração.

Orientador: Cesar Alexandre de Souza

Versão Corrigida
(versão original disponível na Faculdade de Economia, Administração e Contabilidade)

SÃO PAULO
2011

FICHA CATALOGRÁFICA
Elaborada pela Seção de Processamento Técnico do SBD/FEA/USP

Silva, Daniel Bittencourt
Recuperação de informação de comunicados à imprensa / Daniel Bittencourt Silva. -- São Paulo, 2011.
147 p.

Dissertação (Mestrado) – Universidade de São Paulo, 2012.
Orientador: Cesar Alexandre de Souza.

1. Press release 2. Notícia 3. Mercado financeiro – Previsão 4. Redes neurais I. Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade II. Título.

CDD – 070.4

Dedico mais este trabalho aos meus queridos pais,

e à minha querida Alessandra.

AGRADECIMENTOS

Em primeiro lugar quero agradecer à minha namorada Alessandra, mais uma vez, por alimentar meus sonhos, pelo carinho e pela paciência. Sem ela eu não teria conquistado esse objetivo, nem mesmo iniciado essa jornada que hoje tenho orgulho de completar.

À minha família, quero agradecer: ao meu pai Wilson pela compreensão e por fomentar meus planos, à minha mãe pelo carinho e pela torcida, ao meu irmão Rafael pela grande ajuda na correção do trabalho e à minha querida irmã Raquel.

Agradeço ao pessoal da Vector Investimentos por incentivar o mestrado, contribuir com meu trabalho, e pelos recursos disponibilizados. Tem sido muito gratificante trabalhar com colegas bastante qualificados numa cultura empresarial que promove a inovação.

Agradeço ao meu orientador Cesar pela liberdade cedida no desenvolvimento do tema, pela amizade, interesse, e por também apostar na idéia. Aos membros da banca, ao Dr. Christian Zimmer, ao prof. Dr. Eduardo Kayo e ao prof. Dr. Ronaldo Zwicker (Deus o tenha).

Quero também fazer um agradecimento geral a todos os meus professores da FEA e do IME pelos cursos ministrados e pelas dúvidas tiradas tanto pertinentes às disciplinas quanto à minha dissertação.

Study the past, if you would divine the future.

- Confucius

RESUMO

Neste trabalho os retornos de ativos financeiros são modelados com base em dados não-estruturados de notícias aperiódicas. O cerne do tratamento de tais dados está na Recuperação da Informação, que se importa em aplicar uma transformação suficientemente representativa das estruturas de contexto gramatical para o algébrico.

Trabalhos anteriores mostram que não só o timing (GOODHART, HALL, *et al.*, 1993), tipo (ÄIJÖ, 2008), ou contexto no momento de publicação da notícia (BEBER e BRANDT, 2010), mas também o conteúdo léxico (WÜTHRICH, PERMUNETILLEKE, *et al.*, 1998; LAVRENKO, SCHMILL, *et al.*, 2000; SCHUMAKER e CHEN, 2009) explicam movimentos de mercado.

Este trabalho tem por objetivo estender a discussão já fundamentada sobre a semântica das publicações, com a proposta de aprofundar a avaliação da sensibilidade paramétrica do ajuste de modelos já encontrados na literatura. Questões sobre a metodologia de pré-processamento dos textos são exploradas por meio de diversos experimentos.

Por fim, corrobora-se os resultados prévios sobre a possibilidade de aplicar-se uma classificação preditiva sobre movimentos do mercado acionário norte-americano, particularmente dentro da janela de 15 a 25 minutos após a publicação de um *press release*. Contudo, não são encontradas evidências com $\alpha \leq 0{,}01$ de que regressão tenha a mesma capacidade, nem de que exista antecipação pelo mercado da informação a ser anunciada.

ABSTRACT

In this paper the returns of financial assets are modeled based on data from unstructured aperiodic news. The core processing of such data in is information retrieval, which deals with applying a transformation sufficiently representative from structures of grammatical to algebraic contexts.

Previous works show that not only the news timing (GOODHART, HALL, *et al.*, 1993), its type (ÄIJÖ, 2008), or its context in the moment of announcement (BEBER e BRANDT, 2010), but also its lexical content (WÜTHRICH, PERMUNETILLEKE, *et al.*, 1998; LAVRENKO, SCHMILL, *et al.*, 2000; SCHUMAKER e CHEN, 2009) explain movements on the market.

This work aims to extend the ongoing discussion over the semantics of publications, with the proposal of further evaluating the sensitivity of the model's parametric fitting found in the literature. Questions raised about the methodology of pre-processing of texts are explored through various experiments.

Finally, previous results on the possibility of applying a predictive classification of the U.S. stock market movements are corroborated, particularly inside a time frame from 15 to 25 minutes after the publication of a press release. However, no evidence is found with $\alpha \leq 0.01$ that regression has the same capacity, nor that there is anticipation of market information to be announced.

SUMÁRIO

SUMÁRIO xv
LISTA DE ACRÔNIMOS xix
LISTA DE TABELAS xxiii
LISTA DE FIGURAS xxv
LISTA DE SÍMBOLOS xxvii
LISTA DE ALGORITMOS xxix
LISTA DE QUADROS xxxi
CAPÍTULO 1 INTRODUÇÃO 33
1.1 Problema 35
1.2 Objetivos do Trabalho 36
1.3 Justificativa 38
1.4 Hipóteses 39
1.5 Descrição dos Capítulos 40
CAPÍTULO 2 RECUPERAÇÃO DA INFORMAÇÃO 43
2.1 Processamento de Linguagem Natural 44
2.2 Recuperação da Informação 45
2.3 Hipótese da Recuperação de Palavras-Chave 45
2.4 Definições de IR 46
2.5 Estrutura de FIR 47
2.6 Pré-Processamento (Pre-Processing) 48
2.6.1 Extração de Atributos (Feature Extraction) 48
2.6.1.1 Palavras de Parada 49
2.6.1.2 Radiciação 50
2.6.1.2.1 Algoritmo de Porter 52

2.6.1.3 Nota sobre algoritmos para a Língua Portuguesa 53
2.6.2 Seleção de Atributos (Feature Selection) 53
2.6.2.1 B: Boolean Weighting 55
2.6.2.2 TF: Term Frequency 55
2.6.2.3 TF-IDF: Produto TF x IDF 56
2.6.2.4 IG: Information Gain 56
2.6.2.5 MI: Mutual Information 57
2.6.2.6 Poda de Palavras 58
2.6.3 Representação de documentos (Document Representation) 59
2.7 Indução ou Processamento (Induction) 59
CAPÍTULO 3 NOTÍCIAS FINANCEIRAS 61
3.1 Tipos de Notícias 61
3.2 Notícias Corporativas 62
3.2.1 Exemplo descritivo 63
3.2.2 Política de Divulgação 65
3.3 Trabalhos Anteriores em FIR 66
3.4 Hipótese do Mercado Eficiente 74
CAPÍTULO 4 PROCESSAMENTO E INFERÊNCIA 77
4.1 Perceptron Multicamadas 77
4.1.1 Descida do Gradiente 78
4.1.2 Método de Newton 79
4.1.3 Algoritmo de Retropropagação dos Erros 80
4.1.4 Gradiente Conjugado Escalado (SCG) 83
CAPÍTULO 5 METODOLOGIA 85
5.1 Fontes de notícias 85
5.2 Arquitetura do sistema 87
5.2.1 Base de Dados de Notícias 87

5.2.2 Pré-Processamento 90
5.3 Algoritmos de Inferência 95
5.4 Definição do Modelo 96
5.5 Medidas de Desempenho 97
5.5.1 Matriz de Confusão 98
5.5.2 Acurácia 99
5.5.3 Precisão e Revocação 101
5.5.4 MSE: Erro Quadrado Médio 103
5.5.5 Correlação 103
5.5.6 Classificador Aleatório, Solução Trivial e Classes Equiprováveis 104
CAPÍTULO 6 ANÁLISE DOS EXPERIMENTOS 105
6.1 Base de Dados 105
6.2 Primeira Classificação 108
6.3 Sensibilidade aos Principais Parâmetros da MLP 112
6.4 Regressão: Funciona? 118
6.5 Efeito da Dimensão Léxica 121
6.6 Efeito da Pesagem 124
6.7 Efeito do Tempo 126
CAPÍTULO 7 CONCLUSÃO 129
7.1 Pressupostos assumidos 129
7.2 Hipóteses 130
7.3 Discussão Adicional 131
7.4 Conclusões dos Experimentos 132
7.5 Contribuições 133
7.6 Limitações do Trabalho 133
7.7 Sugestão de Trabalhos Futuros 134
CAPÍTULO 8 BIBLIOGRAFIA 137

APÊNDICE A 147

APÊNDICE B 148

LISTA DE ACRÔNIMOS

BP *Backpropagation*: Algoritmo de retro-propagação dos erros

B *Boolean Weighting*

BFGS *Broyden-Fletcher-Goldfarb-Shanno quasi-newton*

C# *C-Sharp*, linguagem de programação OO

CDF *Category Discrimination Frequency*

CEO *Chief Executive Officer*

CGL *Conjugate Gradient with Line Search*

CRR *Company Revenue Relation*: Relação de Retorno da Empresa

CS *Chi-Squared*

CTF *Collection Term Frequency*

CVM Comissão de Valores Mobiliários

DAX30 *Deutscher Aktien IndeX: Índice do Mercado Alemão de Ações*, índice composto pelas 30 maiores empresas negociadas na Bolsa de Valores de Frankfurt

DEM Marco alemão

DET *Detection Error Tradeoff*

DF *Document Frequency*

DJI *Dow Jones Industrial Average*

DMA *Direct Market Access*

DST *Daylight Saving Time*

EBITDA *Earnings Before Interest, Taxes, Depreciation and Amortization*

EMH *Efficient Market Hipotesis*: Hipótese do Mercado Eficiente

EPS *Earnings per Share*

FIR *Financial Information Retrieval*: Recuperação de Informações Financeiras

FTSE-100 *Financial Times Stock Exchange*: Índice composto pelas 100 empresas mais capitalizadas do Reino Unido listadas na LSE

GICS *Global Industry Classification Standard*, taxonomia de indústrias

GMT *Greenwich Mean Time*

HEM Hipótese do Mercado Eficiente

HSI *Hang Seng Index*, índice composto por 45 empresas mais capitalizadas de Hong Kong Stock Exchange

HTML *HyperText Markup Language*, linguagem de marcação para web

Ibovespa *Índice da Bolsa de Valores do Estado de São Paulo*

IDE *Integrated development environment*

IDF *Inverse Document Frequency*

IG *Information Gain*

IPO *Initial Public Offering*

IR *Information Retrieval*: Recuperação da Informação

JPY Iene japonês

KDD *Knowledge Discovery in Databases*

kNN k-Nearest Neighboors

LM Algoritmo de Levemberg-Marquadt

LSE *London Stock Exchange*: Bolsa de Valores de Londres

MI Mutual Information

ML *Machine Learning*

MLP *Multilayer Perceptron*: Perceptron Multicamadas

MSE *Mean Squared Error*: Erro quadrado médio

MSPE *Mean Squared Prediction Error*: Erro quadrado médio de previsão

NaN *Not a Number*

NASDAQ *National Association of Securities Dealers Automated Quotations*

NIKKEI 225 Índice do mercado de ações da TSE

NLP *Natural Language Processing*

NYSE *New York Securities Exchange*

OLS *Ordinary Least Squares*: Regressão Linear

OO Orientação a objetos

OR *Odd's Ratio*

PCA *Principal Component Analysis*

PHP *Hypertext Preprocessor*, linguagem de programação interpretada livre de www

PIB Produto Interno Bruto

RDBMS *Ralational database management system*

RI Relação com Investidores

SCG *Scaled Conjugate Gradient*: Algoritmo de Gradiente Conjugado Escalado

SEC Security Exchange Comission

SQL *Structured Query Language*, linguagem de consulta declarativa para RDBMS

STI *Straits Times Index*, índice composto pelas 30 empresas mais capitalizadas de Singapura

SVM S*upport Vector Machine*

SVR S*upport Vector Regression*

TF *Term Frequency*

TI Tecnologia da Informação

TRAC *Text Reckoning And Compiling*

TS *Term Strenght*

URL *Uniform Resource Locator*, endereço em rede

USD Dólar norte-americano

UTC *Coordinated Universal Time*

UTF-8 *8-bit Unicode Transformation Format*

VBA *Visual Basic for Applications*, linguagem de programação orientada a eventos e associada a IDEs da Microsoft

WDF *Within Document Frequency*

WWW *Worldwide Web*: Rede mundial ou “internet”

LISTA DE TABELAS

Tabela 2.1 Exemplo de Variedade Sucessória 51
Tabela 2.2 Tabela de Contingência para Termos e Categorias 57
Tabela 5.1 Volume de Notícias por Fonte (retirado de THOMAS, 2003) 87
Tabela 5.2 Tabela de contingência de IR 99
Tabela 6.1 Coleções de artigos em estudos 108
Tabela 6.2 Partição do intervalo em Classes 109
Tabela 6.3 Parametrização do problema inicial 110
Tabela 6.4 Desempenho de Classificação 112
Tabela 6.5 Sensibilidade ao número de neurônios ocultos 113
Tabela 6.6 Sensibilidade às falhas de validação 116
Tabela 6.7 Parametrização da regressão 118
Tabela 6.8 Desempenho da regressão por número de neurônios ocultos 118
Tabela 6.9 Desempenho da classificação por número de classes 120
Tabela 6.10 Desempenho da regressão por dimensão léxica 122
Tabela 6.11 Termos removidos entre N=64 a N=32 123
Tabela 6.12 Desempenho de classificação por Método de Pesagem 124
Tabela 6.13 Desempenho de classificação após a publicação 126
Tabela 6.14 Desempenho de classificação prévia à publicação 127

LISTA DE FIGURAS

Figura 2.1 Níveis de informação gramatical 44
Figura 2.2 Esquema geral da metodologia de FIR 47
Figura 2.3 Ilustração da Poda de Dicionário (YANG e PEDERSEN, 1997) 58
Figura 3.1 Exemplo de divulgação dos resultados trimestrais –Nike Inc. 64
Figura 3.2 Preço e volume de NKE antes e após divulgação de resultados 64
Figura 4.1 Perceptron de Rosenblatt 78
Figura 4.2 Arquitetura de uma Feedforward MLP retropropagada simples..... 81
Figura 5.1 Fluxograma do pré-processamento 90
Figura 5.2 Matriz de Contingência genérica 98
Figura 5.3 Gráfico *P* x *R* com 11 pontos de 50 consultas a um sistema da coleção de dados TREC(MANNING, RAGHAVAN e SCHÜTZE, 2009, p. 160). 102
Figura 6.1 Contagem de termos radicados no dicionário universal 105
Figura 6.2 Publicação total de comunicados ao longo da semana..... 106
Figura 6.3 Publicação total de comunicados ao longo do dia 107
Figura 6.4 Instantes de maior publicação 107
Figura 6.5 Matrizes de Confusão de (a) Treinamento (b) Validação (c) Teste e (d) Total 111
Figura 6.6 Número de neurônios ocultos: Curvas de desempenho e Nível descritivo da acurácia 114
Figura 6.7 Número máximo de falhas: Curvas de desempenho e Nível descritivo da acurácia 117
Figura 6.8 Número de classes: Curvas de desempenho e Nível descritivo da acurácia 121
Figura 6.9 Dimensão léxica: Curvas de desempenho e Nível descritivo da acurácia 123
Figura 6.10 Métodos de Pesagem: Curvas de desempenho e Nível descritivo da acurácia ... 125
Figura 6.11 Tempo: Curvas de desempenho e Nível descritivo da acurácia..... 128

LISTA DE SÍMBOLOS

$\mathbb{N}$	conjunto dos números naturais
$\mathbb{R}$	conjunto dos números reais
$M_{m,n}(\mathbb{R})$	conjunto das matrizes reais
$\mathbb{R}^{\mathrm{m,n}}$	matriz real m por n
$x \in X$	elemento x pertence ao conjunto X
α	nível de confiança estatístico
F	matriz de frequências
f_{ij}	frequência da j-ésima lexia do i-ésimo documento
d_i	vetor de frequências do i-ésimo documento
l_j	vetor de frequências da j-ésima lexia
A^T	matriz transposta de A
$*$	produto matricial
$.*$	produto matricial por elemento pareado
$tr(A)$	traço de A
$P(.)$	operador probabilidade
$P(.\|.)$	operador probabilidade condicional
$P_{H_0}(.)$	probabilidade sob hipótese nula
$max(.)$	operador máximo
$\cup$	operador união
$\cap$	operador interseção
$\frac{\partial}{\partial x}$	derivada parcial em relação à x
∇_x	gradiente em relação à x
$\{x_i\}_{i=1}^{n}$	conjunto de x_1 a x_n
$\|\{\ldots\}\|$	cardinal do conjunto
y	variável observada
$\hat{y}$	variável estimada

X	matriz de entrada
$\overline{\text{acur}}$	acurácia observada
$Prec$	precisão média observada
$Revoc$	revocação média observada
$N(\mu, \sigma^2)$	normal com esperança μ e variância σ^2
$B(p)$	bernoulli com probabilidade p de sucesso
$Bin(n, p)$	binomial com n lançamentos e probabilidade p de sucesso
t_n	t-student com n graus de liberdade
$\mathcal{F}(.)$	classificador
$\mathbb{E}(.)$	operador esperança
$\mathbb{V}(.)$	operador variância
$O(.)$	complexidade computacional
ρ	nível descritivo
$\rho_{x,y}$	correlação entre x e y
$r_{x,y}$	correlação observada entre x e y

LISTA DE ALGORITMOS

Pseudocódigo 5-1 Extração de Atributos 91
Pseudocódigo 5-2 Seleção de Atributos 93
Pseudocódigo 5-3 Representação de Documentos 94

xxx

LISTA DE QUADROS

Quadro 5-1 Extrato de código-fonte de um press release na PR Newswire 89
Quadro 5-2 Extrato de Documento armazenado depois da filtragem 89
Quadro 5-3 Extrato do vetor de atributos extraídos 92
Quadro 5-4 Extrato da seleção de atributos 94
Quadro 5-5 Extrato da matriz F de frequencias 95

CAPÍTULO 1 INTRODUÇÃO

A revolução trazida pela *world wide web* na década de 90 mudou completamente o comportamento da sociedade. Primeiro apareceram o e-mail, os *websites* e as ferramentas de busca: artefatos tão profundamente inseridos no cotidiano que podem – em muitos aspectos – parecer indistintos de um telefone ou luz elétrica: muitas pessoas dependem totalmente de tais recursos. Então vieram as compras online, as redes sociais, vídeos digitais, compras coletivas e muitos outros serviços online que trazem cada vez mais a rotina dos indivíduos para dentro da web. E tudo isso vem sendo alimentado por computação móvel, cabos de fibra ótica, provedores mais próximos e servidores mais poderosos.

A revolução não foi menos impactante para o mercado financeiro. Vimos o amadurecimento das transações eletrônicas, internet banking, *homebrokers* e DMA[1], para citar algumas das mudanças. No que tange o assunto deste trabalho, houve a canalização de notícias para o formato digital. O rompimento do paradigma impresso conduziu forçosamente jornais e revistas a irem online. Mais que isso, as notícias foram popularizadas, com o surgimento de *weblogs*, fóruns de discussão e (até mesmo) o Twitter®. Ao encontro com o mercado financeiro, os órgãos de regulamentação passaram a exigir o arquivamento de submissões no formato eletrônico com fácil acesso público[2], a publicação de informações internas da companhia por via de agências de comunicação ao mercado, e as áreas de relacionamento com investidores das empresas passaram a disponibilizar muito mais conteúdo nos sites das empresas.

---

[1] Direct Market Access (Acesso Direto ao mercado) é a comunicação do cliente diretamente com o mercado, através da canalização de suas ordens por meio da infra-estrutura do corretor, o que permite controle total sobre suas emissões de ordens com baixa latência (JOHNSON, 2010).

[2] EDGAR (Electronic Data-Gathering, Analysis, and Retrieval system) é uma base de dados online de acesso público datada de 1996, para coleta, validação e indexação dos relatórios submetidos pelas companhias de acordo com a lei de publicação da U.S.SEC (Securities and Exchange Commission).

Contudo, essa abundância de informação online é subutilizada. Dados numéricos, e.g. cotações históricas, índices macroeconômicos e *order book*, são estudados há muitas décadas, na busca pela identificação de padrões de comportamento do mercado. No entanto, a informação não estruturada do conteúdo textual de notícias, comunicados ao mercado e fóruns de discussão, apesar de vasta, é muito pouco explorada pela academia. Se de um lado há problemas de ruído, ambigüidade, e custo no tratamento de tais dados, do outro há a antecipação de fatos e eventos, o comportamento psicológico dos investidores e a oportunidade de diferenciação num mercado dominado por modelos clássicos.

O estudo de anúncios periódicos, tais como divulgações macroeconômicas e publicação de resultados de empresas, é uma área de pesquisa clássica em economia e finanças. Nessa área as abordagens clássicas costumam ser o estudo do evento de publicação (JENNINGS e STARKS, 1986), dos números divulgados (ALVARO ALMEIDA, 1998) ou da diferença do resultado em relação à expectativa do mercado (LOUHICHI, 2008).

Mas a área conhecida como Recuperação de Informação ("IR") trabalha com outra abordagem que não o estudo de eventos. IR é a ciência da pesquisa por documentos e pela informação contida neles, em seus meta-dados, em bancos de dados ou na *worldwide web* (adaptado de GERARD SALTON, 1986). Mais recentemente, um nicho de pesquisa de IR vem estudando a relação de documentos financeiros com o mercado, tais como notícias, comunicados, resultados trimestrais e fóruns de discussão. Chamaremos este novo e crescente nicho de Recuperação de Informações Financeiras ou "FIR" (para *Financial Information Retrieval*) deste ponto em diante.

A área de IR sofreu uma guinada no final dos anos 90 com a explosão do conteúdo informativo na www descrita acima. Com o crescimento da rede, novas pesquisas na área e o declínio do custo de hardware, a demanda por ferramentas de IR vem sendo mais bem suprida (Manning et.al., 2009). O reflexo dessas mudanças foi exibido pela primeira na área de FIR vez com a tese de Peramunetilleke (1997) intitulada "*A system for Exchange Rate Forecasting Using News Headlines*" e orientada pelo Dr. Wüthrich, um dos pioneiros do assunto.

Ainda são muitas as dificuldades encontradas pela FIR na extração de dados a partir de notícias, principalmente no que diz respeito à semântica gramatical, como Kroha e Baeza-Yates (2004) apontam. No entanto vem se notando muitos resultados positivos e desenvolvendo-se a pesquisa no assunto.

Para citar alguns, Wüthrich, Peramunetilleke (1998) demonstraram que o uso de dados textuais pode levar a uma classificação correta dos movimentos de mercados ao fim do dia; Lavrenko e Schmill (2000) e Mittermayer (2004) afirmam que seus sistemas de recomendação de notícias relevantes e previsão têm um retorno médio positivo; Peramunetilleke e Wong (2000) comparam o resultado de um mesmo sistema de previsão que utiliza apenas entradas numéricas com outro que use também entradas textuais e verificam um desempenho de previsão superior ao segundo; e Thomas (2003) confirmou a correlação positiva entre o volume de informação veiculada e o volume negociado.

1.1 Problema

A pesquisa econômica tem uma extensa lista de publicações envolvendo notícias. Os objetos de estudo são em geral os números macroeconômicos de um país (e.g. produto interno, inflação, desemprego) ou os principais eventos corporativos (e.g. dividendos, resultados trimestrais e anuais). Os estudos costumam envolver três metodologias centrais na manipulação destas notícias: o momento da publicação no tempo (GAJEWSKI, 1999; JENNINGS e STARKS, 1986); o número do objeto publicado[3] (ALMEIDA, GOODHART e PAYNE, 1998); ou a diferença entre a expectativa de mercado e o número de fato publicado (BEBER e BRANDT, 2010; LOUHICHI, 2008).

Como visto, modelos financeiros de gestão de investimentos baseados em dados quantitativos têm sido extensamente desenvolvidos (BASEL M. A. AWARTANI, 2005) (PRATYOOSH RAI, 2011), mas relativamente pouca atenção tem sido dada à grande quantidade de informação textual disponibilizada pelas mídias. Em função da grande dificuldade de extrair informação relevante de dados não-estruturados, i.e., não numéricos, as publicações que se utilizam dessa abordagem desfrutam de menos popularidade que as versões com mineração de dados estruturados (MITTERMAYER, 2004). Por outro lado, como apontam Wüthrich e Peramunetilleke (1998), declarações textuais, tais como comunicados à imprensa (*press releases*), notícias ou relatórios contém não só o efeito como a possível causa de um movimento de mercado (mais detalhes em 2.1 Processamento de Linguagem Natural).

[3] I.e. projeção de crescimento do PIB, inflação do mês, taxa de desemprego corrente, respectivamente aos exemplos dados acima.

É de se esperar, portanto, que a análise mais profunda de notícias possa incorporar vantagem informacional sobre modelos puramente numéricos, que em certa medida, seriam capazes de detectar apenas mudanças reflexivas no mercado. Sistemas de interpretação textual, por outro lado, podem aproveitar o fato relevante gerador do movimento de mercado e generalizar contextos gramaticais que identifiquem oportunidades de *trade* antes que isso se reflita sobre as variáveis numéricas de preço, volatilidade ou *spread*.

Da mesma forma, podem atuar de maneira preventiva para gestão de risco, monitorando notícias relativas a eventos de risco, e buscando notícias que contenham palavras-chave importantes para um determinado ativo sob gestão. Não apenas modelos de gestão de carteira ou risco, mas também modelos macroeconômicos poderiam se beneficiar com a captura de informação textual, absorvendo informação e texto sobre notícias gerais, internacionais e macroeconômicas, monitorando composições léxicas que reflitam a expectativa geral sobre economia, de acordo com palavras que reflitam esse humor ou sentimento.

A análise de textos incorpora uma nova dimensão à tomada de decisão de investimentos. Além de usar puramente textos, como é o foco desse trabalho, pode-se usar em paralelo com outros modelos numéricos puros, ou em conjunto com modelos híbridos, concluindo Thomas e Sycara (2000) ser esta a melhor alternativa.

A representação de um *press release* é simplificada com a quebra da estrutura gramatical do texto em palavras e subseqüente representação vetorial da freqüência das palavras no texto. No interior de uma notícia, esperamos encontrar palavras que remetam ao sujeito das orações sobre o qual a informação trata.

1.2 Objetivos do Trabalho

O objetivo central deste trabalho é avaliar a capacidade de previsão dos *press releases* em relação a um ativo financeiro sujeito. Especificamente, o problema se decompõe nos seguintes objetivos secundários:

1. Desenvolver um sistema especialista de FIR com base na literatura prévia sobre o tema e documentá-lo;
2. Verificar a capacidade das palavras fora de contexto semântico (ver 2.1 - Processamento de Linguagem Natural) de explicar os movimentos futuros de mercado, como relatado nas pesquisas anteriores;
3. Testar a sensibilidade dos resultados de previsão aos parâmetros da rede neural;
4. Comparar os resultados de dois diferentes métodos de inferência: regressão e classificação;
5. Explorar os resultados frente às diversas parametrizações que envolvem a metodologia clássica de FIR;
6. Verificar os resultados de previsão no tempo, em torno do instante de publicação da notícia.

Os objetivos se estendem longitudinalmente sobre a metodologia empregada em FIR, abordando as diversas etapas que envolvem a pesquisa no assunto. Começa com o desenvolvimento de um sistema especialista para captura e armazenamento de notícias, seguido pela sensibilidade do sistema de pré-processamento e finalmente na inferência sobre os dados filtrados no tempo. Toda essa trajetória envolve diversas possibilidades sobre como abordar o problema, e o trabalho tem o objetivo de documentar cada um desses resultados.

Considerando-se que a metodologia ora empregada envolve diversas etapas encadeadas com uma vasta combinação de parâmetros e métodos, e que há uma falta de relatórios de sensibilidade dos modelos desenvolvidos na literatura, existe a chance de que os resultados possam variar muito, ou seja, de que a metodologia não seja resistente a mudanças sutis na parametrização do modelo. Há até mesmo a chance de que os resultados positivos sejam intervalos estreitos de ótimo local, o que representa um risco ainda maior para a metodologia.

Assim, este trabalho também pretende documentar os resultados de experimentos sobre secções transversais de parâmetros com o objetivo de entregar uma maior compreensão sobre as características do modelo. A compreensão é entendida como a importância de cada parâmetro para o sucesso na execução da tarefa e qual a sua margem possível de mudança. Por exemplo, se o número de neurônios ocultos numa rede neural pode ser arbitrariamente grande, qual o intervalo ótimo da dimensão léxica e qual a vantagem de usar uma pesagem de

termos com *Mutual Information* sobre *Information Gain*, são exemplos de questões examinada neste trabalho.

Além disso, um aspecto muito importante é o impacto das notícias ao longo do tempo. Na pesquisa econômica clássica é muito comum ver a aplicação de modelos de séries temporais para avaliar o impacto de uma notícia positiva ("good news") no preço do ativo ao longo do tempo. Já foram encontradas evidencias de uma noticia ter uma janela de não mais que 10 minutos (JAMES M. PATELL, 1984) a até 2 horas (GOODHART, HALL, *et al.*, 1993), para que sua informação seja completamente absorvida pelo mercado.

1.3 Justificativa

Até o momento de elaboração deste trabalho, as pesquisas focaram bastante sobre o procedimento de inferência, isto é, sobre os resultados obtidos com diferentes tipos de algoritmos de classificação, porém ainda não há resultados suficientes de testes mais básicos sobre o pré-processamento. O único trabalho que chega mais próximo desse objetivo é o de Mittermayer e Knolmayer (2006), no qual se discute o impacto de diferentes métodos de realizar um tratamento prévio dos textos. Desta forma há carência por uma maior exploração do espaço paramétrico que a metodologia envolve.

Poucos resultados de significância estatística dos resultados foram reportados, além daqueles que podem ser inferidos diretamente sobre os números finais. Contudo, uma documentação mais detalhada dos testes é pretendida pelo presente trabalho.

A importância do aprofundamento das pesquisas no campo de FIR está imediatamente ligada à gestão de investimentos, com o detalhamento de ferramentas que auxiliem a tomada de decisão do gestor de carteiras. Para controle de risco, a metodologia apresentada pode fornecer meios de se extrair automaticamente notícias com maior probabilidade de impactar o mercado de forma severa, a partir de um volume muito grande de informação. E do lado de regulação do mercado, é possível detectar assimetria informacional tal como *insider trade*s, ou em outras palavras, vazamento de informação privilegiada.

1.4 Hipóteses

As hipóteses do trabalho iniciam pela abordagem com IR, que em sua raiz assume a Hipótese da Recuperação da Informação, de acordo com a qual se um documento possui uma palavra-chave, então o documento é sobre essa palavra e a palavra é sobre o documento.

Trazendo o conceito para finanças, pela mesma hipótese pode-se afirmar que se uma notícia contém referência ao nome de uma empresa (ou ao seu *ticker* – código identificador de um ativo financeiro), então a notícia é sobre a empresa e a empresa é sobre a notícia. Caso contrário, as notícias são totalmente aleatórias, e não é possível inferir nada acerca de seu conteúdo textual. Uma possível interpretação para a hipótese no contexto financeiro, dada uma menção para a $empresa_j$ e uma $notícia_i$ seria

$$P\left(empresa_j, notícia_i\right) > P\left(empresa_j\right)P(notícia_i).$$

Nesta pesquisa parte-se do pressuposto de que Hipótese da Recuperação da Informação seja verdadeira, ou seja, de que as notícias e as palavras carregam uma relação probabilística entre elas.

O objetivo primordial deste trabalho é testar a metodologia já empregada em FIR e validar ou rejeitar a possibilidade de prever o movimento de ativos financeiros com base absolutamente em conteúdo textual de notícias.

Contudo em finanças, a discussão sobre previsão de séries financeiras se estende sobre a Hipótese da Eficiência do Mercado, pela qual o preço de um ativo reflete toda a informação disponível sobre ele até o dado momento. No capítulo 3.4 (Hipótese do Mercado Eficiente) o assunto é discutido com maior detalhe. Pela hipótese, não é possível prever o movimento futuro de um ativo e auferir-se lucro com isso, pois a competição entre os participantes é eficiente suficiente para que a margem de lucro sobre a previsão seja nula e o ajuste à chegada de nova informação imediata.

Desta forma, definimos as seguintes hipóteses para este trabalho:

$H_{0,1}$: O conteúdo textual de comunicados ao mercado não explica o movimento futuro de nenhum ativo financeiro.

$H_{a,1}$: O conteúdo textual de comunicados ao mercado explica o movimento futuro de algum ativo financeiro.

$H_{0,2}$: O mercado não antecipa a informação a ser publicada em uma notícia.

$H_{a,2}$: O mercado antecipa a informação a ser publicada em uma notícia.

Na primeira hipótese, não é possível testar empiricamente se o conteúdo textual pode explicar o movimento de qualquer ativo. Ademais, é importante notar que se $H_{0,1}$ for rejeitada, então $H_{0,2}$ é conseqüentemente rejeitada também, por conter $H_{0,1}$. Porém o inverso não é verdadeiro.

A questão da antecipação é algo que envolve a regulação do mercado, mas também sua eficiência: se é possível detectar que uma informação privada vaza antes de ser publicamente anunciada, então é um indício para os problemas de Assimetria da Informação no mercado e também do Principal-Agente (que não discutiremos aqui).

1.5 Descrição dos Capítulos

Os capítulos 2, 3 e 4 tratam da revisão teórica do trabalho, que por sua natureza interdisciplinar, foi dividido em Recuperação da Informação (IR), Recuperação de Informação Financeira (FIR), e Inferência, respectivamente. O capítulo 5 trata da Metodologia desta pesquisa, e o capítulo 6 analisa os resultados com os métodos empregados. O capítulo 7 conclui o trabalho.

No capitulo 2 será dada uma introdução à Recuperação da Informação, seu jargão e metodologia quando aplicada à classificação automática de documentos. As subseções em 2.5 descrevem as 3 etapas fundamentais de tal disciplina: Extração de Atributos, Seleção de Atributos e Representação dos Documentos.

O capítulo 3 discute os tipos de notícias financeiras, suas fontes, e a política de divulgação envolvendo comunicados à imprensa. Mas faz principalmente uma revisão teórica dos trabalhos em FIR.

O capítulo 4 faz uma introdução às redes neurais com um breve histórico deste tópico e da formulação matemática do algoritmo de decida do gradiente, e sua aplicação à aprendizagem supervisionada da rede por retropropagação dos erros. As redes neurais são usadas neste trabalho para classificar as notícias e regredir os movimentos dos ativos.

O capítulo 5 descreve a montagem das bases de dados, arquitetura do sistema, algoritmos de pré-processamento de textos, algoritmos de inferência e medidas de desempenho com testes de hipótese.

O capítulo 6 se inicia com um breve exemplo de notícia retirada da base de dados e analisa seu impacto sobre o mercado. Depois descreve os resultados sobre a tarefa de classificação empregando a mesma metodologia encontrada previamente na literatura, e a sensibilidade dos resultados quanto ao número de neurônios ocultos da rede neural e critério de validação cruzada. O mesmo teste de sensibilidade é então conduzido com uma rede para regressão.

O capítulo segue com experimentos sobre os efeitos da dimensão léxica, pesagem de termos e tempo, que são parâmetros fundamentais na montagem de um sistema de FIR.

A conclusão do trabalho é feita no capítulo 7, com discussão das hipóteses formuladas, explicitação das limitações e sugestões para pesquisa futura.

CAPÍTULO 2 RECUPERAÇÃO DA INFORMAÇÃO

Esta seção tem como objetivo rever aspectos teóricos importantes que vão ao encontro com a pesquisa em Recuperação de Informações Financeiras ("FIR"). Como já mencionado anteriormente, FIR é definida nesse trabalho como um campo de pesquisa recente gerado pela Recuperação de Informação ("IR"). A pesquisa no campo tem como objetivo filtrar notícias financeiras e extrair padrões textuais delas que representem informação relevante ao estudo de um determinado objeto financeiro. A utilidade final está no suporte à decisão de investimentos, controle de risco e regulamentação do mercado.

Por ser uma área multidisciplinar, as áreas de pesquisa mais influentes serão revistas, nos tópicos que tangenciaram a pesquisa feita até o momento. Pode-se afirmar a partir delas que dois são os pilares fundamentais da pesquisa em FIR – a Recuperação da Informação e a Aprendizagem de Máquina.

Sendo assim, na primeira seção será abordada a pesquisa em Recuperação da Informação (IR), com uma revisão "transversal" das principais definições e metodologias de mineração de textos empregadas na área. Nos capítulos subsequentes a Aprendizagem de Máquina ("ML") será revista, com atenção especial às máquinas de inferência mais utilizadas em FIR. Como síntese das seções 1 e 2, teremos na terceira seção uma discussão ampla e detalhada sobre a pesquisa já feita em FIR, abordada trabalho a trabalho em ordem cronológica. Para concluir, a Hipótese do Mercado Eficiente ("EMH") será brevemente abordada, por ser um tema que tangencia pesquisas envolvendo previsão de séries financeiras.

2.1 Processamento de Linguagem Natural

A literatura de Processamento de Linguagem Natural (NLP) costuma estratificar o conteúdo informacional de um texto em diversas camadas, como exposto na Figura 2.1. Um morfema é a menor unidade gramatical identificável, e compõe a estrutura de uma lexia[4]. Os morfemas mais comuns são: prefixo, radical (ou raiz) e sufixo.

As análises léxica (sobre palavras), sintática (sobre estrutura de frases) e semântica (sobre o significado direto da frase) são muito comuns em computação na construção de compiladores. Um compilador utiliza-as para checar por erros no código e para processar o código escrito. Como exemplo, para um interpretador de aritmética simples, a expressão "3+4" é válida, mas "#+4", "3+4+" não são léxica nem sintaticamente válidas, respectivamente. A análise semântica avalia "3+4" em "7".

A análise pragmática está um nível acima da semântica, uma vez que a frase "Seu cabelo está bonito! Cortou no pet-shop?" é semanticamente positiva para um interlocutor não-pragmático.

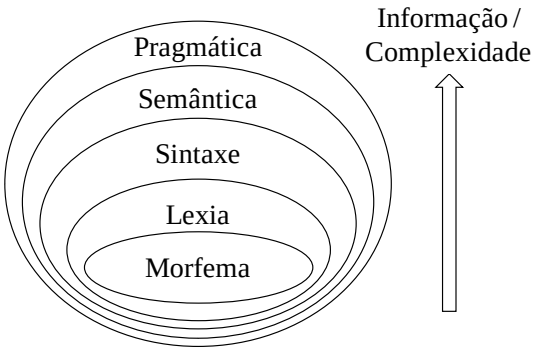

Figura 2.1 Níveis de informação gramatical

Assim, o conteúdo informacional de um texto pode ser decomposto em camadas de complexidade e completude da informação. Nesta pesquisa usaremos um dos níveis mais baixos desse extrato: o nível lexia / morfema. De fato, as palavras (lexias) serão podadas para que apenas o morfema da raiz remanesça. Espera-se que mesmo com tal poda a maior parte do conteúdo informacional permaneça.

[4] Lexia é sinônimo de palavra

Muitos autores consideram a Recuperação da Informação como uma subárea de NLP. A seguir vamos explorar os detalhes dessa metodologia que será empregada extensivamente adiante.

2.2 Recuperação da Informação

Os problemas de indexação e recuperação da informação (IR), também denominada *Knowledge Discovery in Databases* ou KDD (FALINOUSS, 2007), cresceram em importância acompanhando à medida que os dados cresceram em volume. Contudo, não foi até meados do século passado que a área foi identificada, aparentemente nomeada em 1950 por Calvin Mooers, cientista da computação e criador da linguagem de programação TRAC (*Text Reckoning And Compiling*), puramente baseada em textos e orientada a expressões, no início dos anos 60.

Um processo de recuperação da informação se inicia com a entrada de uma consulta ao sistema. O sistema retorna à consulta uma série de objetos (livros, periódicos, imagens, etc.) indexados por sua relevância. Por sua vez, a relevância de um objeto é uma medida de desempenho que reflete a expectativa de adequação deste com a consulta realizada, baseada num critério numérico. Em contrapartida, na FIR a consulta é dirigida ao conteúdo textual das publicações, porém é livre de interface com o usuário.

2.3 Hipótese da Recuperação de Palavras-Chave

Um ponto essencial discutido por Arampatzis e Tsoris (1998) sobre IR é que todo o campo está baseado na assunção de que "*se uma consulta e um documento ambos têm uma palavra-chave em comum, então o documento é sobre a consulta*". Tal argumento pressuposto é identificado como *Naïve Keyword Hypothesis*. Mesmo aplicações mais elaboradas, livres de palavras-chave redundantes ou ruidosas e com ponderação da relevância dos termos, precisam assumir a Hipótese da Recuperação de Palavras-Chave (*Keyword Retrieval Hypothesis*), definida como:

Definição 2.1 (Hipótese da Recuperação de Palavras-Chave):

Uma palavra-chave de um documento determina parte do significado do documento como um todo.

A palavra-chave pode ser uma palavra ou sua raiz, determinada automaticamente por métodos de inferência ou manualmente com intervenção humana.

A hipótese não somente tem sido validada em diversos estudos, como com o desenvolvimento da pesquisa na área atingem-se pontos de confiança cada vez mais distantes e com uma quantidade menor de documentos-exemplos para tomada de decisão (FELDMAN e SANGER, 2007).

2.4 Definições de IR

Antes de seguir com o detalhamento de cada módulo, é importante fazer as seguintes definições iniciais de conceitos relacionados a IR:

- Documento: um documento é o menor objeto textual analisado, por exemplo artigos provindos de um periódico, notícias contidas em um ou mais jornais, páginas de um site, etc.
- Corpus: é o conjunto de todos os documentos disponíveis na base de dados.
- Termo / Lexia: um termo ou lexia é a menor unidade léxica analisada. Um termo pode consistir de apenas uma palavra, uma raiz de palavras ou uma seqüência n-ária de palavras, tal qual uma frase. Na literatura revista, usa-se a denominação term (termo), mas lexia seria a denominação mais apropriada de fato. Neste trabalho, ambas as denominações serão usadas com o mesmo significado apresentado acima.
- Léxico: é o conjunto de palavras, mais precisamente, de lexias. O número de termos distintos que compõe o léxico perfaz a dimensão léxica, importante na definição da topologia das máquinas de aprendizagem na etapa de indução.
- Sintaxe: seguindo a definição dada por Machado, Correia e Capella (2004), sintaxe é o ramo lingüístico que estuda as combinações e relações entre as palavras dentro da oração e das orações dentro do período com a especificação das estruturas internas das mesmas. Tal como o léxico, e relativamente

contrária à semântica, a sintaxe é absolutamente dependente da linguagem comunicada.

- Semântica: para Hurford e Heasley (2004), semântica é o estudo da teoria do significado na linguagem. Em Müller e Negrão (2003), os autores afirmam que a semântica não tem como funcionar a não ser sobre expressões já sintaticamente analisadas, desta forma, sintaxe e semântica não são autônomas, mas absolutamente interdependentes.

2.5 Estrutura de FIR

Com algumas adaptações sobre artigo de Apté, Damerau e Weiss (1994), ao contexto da FIR, que é foco deste trabalho, é possível generalizar a metodologia em 3 módulos seriais após a coleta de notícias:

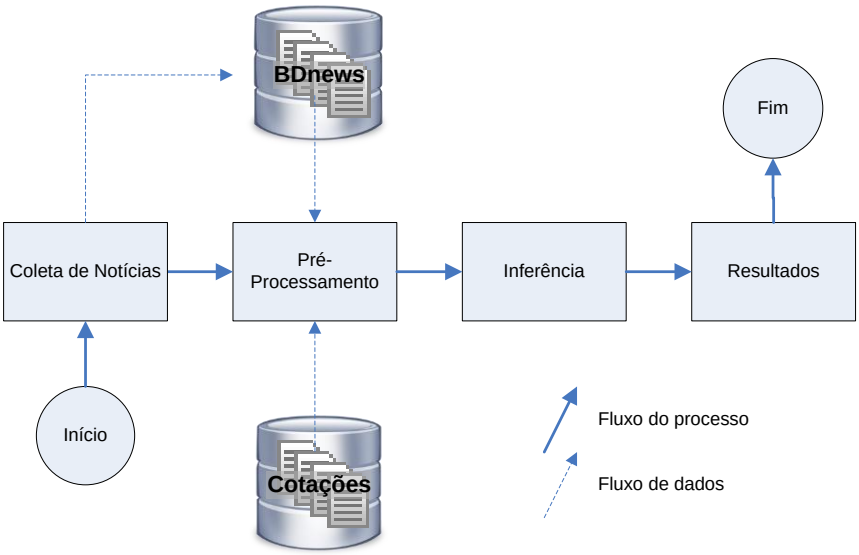

Figura 2.2 Esquema geral da metodologia de FIR

Uma vez especificadas as fontes e/ou bancos de dados, o primeiro módulo é o pré-processamento, que se encarrega de realizar a transformação do conteúdo textual em estruturas de dados numéricas, e é fundamental no sucesso de um projeto de IR(CHO, WÜTHRICH e ZHANG, 1999). O segundo módulo consiste em processar os dados

numéricos utilizando ferramentas de indução, e o terceiro em aferir os resultados obtidos para uma eventual otimização do modelo.

2.6 Pré-Processamento (Pre-Processing)

O pré-processamento dos textos é a essência de IR, por ter a dupla finalidade de transformar o conjunto de dados textuais em dados numéricos e filtrar grande parte do ruído de entrada.

A transformação permite a manipulação algébrica do conteúdo textual. Desfruta de uma gama muito ampla de possibilidades. Contudo, o método mais trivial e canônico, porém robusto (WÜTHRICH, PERMUNETILLEKE, *et al.*, 1998) é a contagem de ocorrências de palavras-chave no documento.

Os filtros aplicados nesta etapa são de granularidade maior, para remoção de grande parte do conteúdo ruidoso dos textos. São distintos dos filtros usados na etapa de processamento alimentados com dados numéricos já transformados.

O módulo de pré-processamento subsume 3 etapas seqüenciais (BRÜCHER, KNOLMAYER e MITTERMAYER, 2002; SOLKA, 2008): Extração de Atributos, Seleção de Atributos e Representação de Documentos. A extração de atributos reúne todas as palavras-chave dos documentos e a seleção de atributos reduz o tamanho do conjunto. A etapa de representação modula a entrega dos dados pré-processados à máquina de aprendizagem.

2.6.1 Extração de Atributos (Feature Extraction)

O objetivo principal da extração de atributos é a geração de uma estrutura gramatical capaz de descrever os documentos suficientemente livre de ruído (adaptado de Brücher, Knolmayer, Mittermayer). Como dito anteriormente, a estrutura gramatical mais simples e largamente usada é uma listagem de lexias (i.e. palavras), mas pode ser uma matriz de correspondências entre termos (SOLKA, 2008), árvores morfológicas ou caracteres (FELDMAN e SANGER, 2007, p. 218-219) ou até mesmo redes semânticas (MA, SHENG e PANT, 2009). De qualquer modo, um atributo (*feature*) em IR equivale a uma unidade da

estrutura gramatical proposta, e em termos estatísticos, corresponde a uma variável livre do modelo.

A geração de estrutura gramatical representativa e suficientemente livre de ruído se traduz em termos computacionais à redução da necessidade de recursos. Consideremos a listagem de termos como forma de ilustração. A primeira etapa da extração de atributos seria coletar todos os termos encontrados nos documentos, gerando assim um dicionário preliminar. Com isso é possível definir a dimensão léxica como o cardinal do conjunto de termos desse dicionário.

Como a dimensão léxica tende a ser extremamente grande – entre centenas de milhares a milhões de palavras para projetos de porte médio (MANNING, RAGHAVAN e SCHÜTZE, 2009) – é interessante filtrá-la a fim de remover palavras que gerem ruído (*stop words*) e reduzi-la para melhorar o desempenho computacional. Com isso chegamos à segunda etapa, na qual é comum aplicar a remoção das palavras e a radiciação dos termos (*stemming*).

É importante notar que a aplicação de quaisquer dessas atividades é facultativa, e faz parte do escopo de um projeto de *text mining*, por ser esta uma questão de pesquisa ainda em aberto (SOLKA, 2008).

A seguir os métodos de mais comuns de remoção de palavras e radiciação serão descritos com maior detalhe.

2.6.1.1 Palavras de Parada

O algoritmo trivial de remoção envolve a pré-especificação das palavras de parada numa lista, para todos os documentos ou dependente do contexto (Solka, 2008). As Palavras de Parada são aquelas que aparecem com freqüência nos textos, mas não adicionam nenhum poder de discriminação à análise de um texto, tais como artigos (‘o’, ‘as’, ‘umas’, ‘todo’, ‘ambas’, etc.), preposições (‘para’, ‘em’, ‘sob’, ‘com’, etc.), conjunções (‘e’, ‘mas’, ‘quanto’, ‘nem’, ‘como, ‘onde’, etc.) pronomes (‘eu’, ‘nós’, ‘se’, ‘lhe’, etc.), numerais (‘1’, ‘dois’, ‘3º’, ‘sétimo’, ‘meio’, etc.), e outras palavras recorrentes nos documentos que o pesquisador julgar indiferente. O objetivo da remoção das palavras de parada é reduzir o ruído provocado pelos termos com poder de explicação menor. Definiremos o poder de explicação ou discriminação de uma lexia como a capacidade dela de caracterizar um padrão no sistema analisado. Uma

vez removidas as palavras de parada, o dicionário está “livre de ruído” (MITTERMAYER, 2004).

2.6.1.2 Radiciação

A radiciação é a redução de variações de uma mesma palavra a uma representação única: o radical (ou estema) (OLIVEIRA, 2010). Brevemente, é a remoção de afixos das palavras, isto é, prefixos e sufixos, fazendo palavras muito próximas assumirem uma mesma característica (*feature*) gramatical.

Um exemplo disso é a radiciação das palavras ‘interessante’, ‘interessado’, ‘interessamos’ e ‘desinteressou’, que para todas gera o mesmo estema ‘interess’, porém só a última possui o afixo ‘des’ que opõe o significado da palavra. Por isso os algoritmos de radiciação eliminam em geral apenas os sufixos. Alguns dos objetivos da radiciação são a simplificação do modelo, a redução da dimensão léxica e, com isso, a redução do custo computacional.

O algoritmo trivial é de procura da palavra numa tabela de estemas. Apesar de ser um algoritmo rápido, pois requer apenas uma busca na tabela para cada palavra dada ($nlogm$), não há uma tabela pronta para o inglês (ZHE e CHIA-HAO, N/D), nem mesmo para o português. Por isso demandaria muito trabalho, já que qualquer língua ocidental tem centenas de milhares de palavras, cada uma com dezenas de flexões e desinências. Da mesma forma o custo de armazenamento em memória RAM de tal tabela para uma língua completa seria considerável, exigindo trabalho com banco de dados ao invés de um simples carregamento de matriz.

Vários algoritmos alternativos atacam esse problema transferindo a carga de memória para o processamento. Um exemplo disso é a Variedade Sucessória, que para cada palavra de um dado corpus, determina-se a distribuição de fonemas de acordo com o número de proferições seguintes. O exemplo da Tabela 2.1 (adaptado de Zhe, Chia-Hao) ilustra o algoritmo para a palavra de teste VESTIDO:

Dado o corpus {ABACATE, IDO, LER, SENTIR, TESTE, VALSA, VER, VESTE, VESTIR, VESTUÁRIO, VOANDO}, podemos contar o número de possibilidade para completar cada raiz de VESTIDO (Tabela 2.1):

Tabela 2.1 Exemplo de Variedade Sucessória

Prefixo	Variedade Sucessória	Letras
V	3	A, E, O
VE	2	R, S
VES	1	T
VEST	3	E, I, U
VESTI	2	D, R
VESTID	1	O
VESTIDO	0	-

A probabilidade de tomar-se uma decisão errada de radiciação pode ser reduzida cortando-se a raiz nos pontos em que um nível de variedade sucessória pré-fixado é atingido, ou nos pontos em que a freqüência volta a subir, como é o caso da raiz 'VEST'.

Tal como o uso grosseiro de probabilidade na Variedade Sucessória, o método de N-Gramas usa algo semelhante à auto-correlação serial para determinar palavras semelhantes. A medida de associação é calculada entre pares de termos baseadas em n-gramas distintos (em geral, digramas). Exemplo com os termos 'Interesse' e 'Desinteressou':

Palavra A: Interessante => `in nt te er re es ss sa an nt te`

Bigramas distintos de A = `in nt te er re es ss sa an`, $|A| = 9$

Palavra B: Desinteressou => `de es si in nt te er re es ss so ou`

Bigramas distintos de B = `de es si in nt te er re ss so ou`, $|B| = 11$

Bigramas comuns entra A e B = `in nt te er re es ss`, $|C| = 7$

Coeficiente de Dice (Similaridade): $S = \frac{2|C|}{|A| + |B|} = \frac{14}{20} = 0{,}70$

O coeficiente pode ser usado com um limiar, acima do qual as palavras são agrupadas (no exemplo acima se espera que o limiar seja maior que 0,7), ou numa matriz de similaridade $n \, x \, n$.

O algoritmo de radiciação mais usado e tradicional é o algoritmo de Porter, que apresenta boa confiabilidade e excelente desempenho e será usado neste trabalho.

2.6.1.2.1 Algoritmo de Porter

O algoritmo de Porter não tem uma abordagem linguística. Baseia-se apenas na observação que o comprimento de uma palavra medido em termos de vogais e consoantes poderia ser uma boa *proxy* para a estemização. A seguinte função resume a essência do algoritmo:

$$[C](VC)^m[V]$$

Onde *C* equivale a uma seqüência de consoantes, *V* a uma sequência de vogais e *m* é a medida do estema. Temos os seguintes exemplos:

Medida	Exemplos
m = 0	PS, OI, ME, CHÁ, BBOOO
m = 1	TEM, UVA, RODO, REGRA
m = 2	COMITE, ARVORE, PRIVADO

Após a medição do estema, algumas regras do tipo *IF-THEN-ELSE* são aplicadas sobre as palavras. As regras são agrupadas em 5 passos para extraírem gradativamente os afixos. Um exemplo de cada passo foi extraído do artigo original com algoritmo para a língua inglesa (PORTER, 1980):

Passo 1(b):	ING →	motoring → motor
Passo 2: ($m > 0$)	ATIONAL → ATE	relational → relate
Passo 3: ($m > 0$)	ICATE → IC	triplicate → triplic
Passo 4: ($m > 1$)	AL →	revival → reviv
Passo 5(a) ($m > 1$)	E →	probate → probat

O algoritmo foi projetado para evitar a sobre-radiciação de palavras curtas. Está disponível em diversas línguas, incluindo o Português (Vieira e Virgil, 2007) e linguagens de programação na página do autor (http://tartarus.org/~martin/PorterStemmer/).

2.6.1.3 Nota sobre algoritmos para a Língua Portuguesa

Dentre os algoritmos de radiciação para língua portuguesa, temos Porter-PT, Orengo, Pegastemming (Vieira e Virgil, 2007), RSLP, STEMBR e Savoy (Flores, 2009). De acordo com Flores, 2009, o algoritmo de Savoy se destaca pela maior precisão e revogação – porém não muito distante dos demais – e pelo menor tempo de execução (speed-ups de 50x Porter e 5x RSLP). O RSLP, porém se mostrou mais confiável, com menor erro de remoção insuficiente de afixos *(understemming*) e menor erro de truncamento.

2.6.2 Seleção de Atributos (Feature Selection)

É esperado que, mesmo com a radiciação e filtragem, o número de termos do dicionário ainda seja muito elevado. O excesso de atributos não só causa lentidão no processo de treinamento, bem como decremento da qualidade do conhecimento extraído (*overfitting*) (NOGUEIRA, 2009).

Desta forma, a seleção de atributos é o processo de redução da alta dimensionalidade do espaço de dados presente para um espaço de atributos com menor dimensão (HAYKIN, 2001, p. 433).

Existem dois principais métodos de Seleção de Atributos: *Bag of Words* e Espaço Vetorial.

O *Bag of Words* merece atenção por ser o modelo trivial na elaboração de um sistema de FIR. Consiste em pré-selecionar um conjunto estático de palavras que um especialista no assunto considere de alta relevância para o assunto estudado. Um exemplo no campo é dado por (WÜTHRICH, PERMUNETILLEKE, *et al.*, 1998), que utilizaram uma seleção com 392 palavras-chave.

Os autores não especificam quais palavras foram usadas, mas para FIR pode-se especular sobre o uso de palavras referentes às divulgações (*earnings*, *quarterly*, *report*, *dividends*, *announces*, *warning*, *call*), mercado (*stock*, *price*, *options*, *bonds*), economia (*payrolls*, *product*, *inflation*, *exchange*), análise de investimento (*EPS*, *profit*, *loss*, *EBITDA*), verbos (*raise*, *fall*, *gain*, *lose*), adjetivos (*bullish*, *bearish*, *up*, *down*), indústria, players, *tickers*, nomes de empresas, classificações de risco, etc.

As desvantagens do *Bag of Words* estão na elaboração manual do dicionário de palavras, na sua natureza estática e na necessidade do apoio de um especialista no assunto, o que dependendo da aplicação em IR é inviável dado o grande número de assuntos que a base de documentos pode conter.

O emprego de métodos dinâmicos de pesagem visa contornar esses obstáculos aplicando heurísticas (validadas empiricamente) sobre todo o conjunto de palavras da base e selecionando as que constituirão o dicionário final. Primeiramente é feita uma classificação dos termos de acordo com o peso para uma determinada métrica (e.g. TFIDF ou IG), e então os termos com maior peso são filtradas (os de peso menor são descartados, ou "podados").

No método de espaço vetorial, a seleção de atributos consiste primeiro na montagem de um dicionário, que equivale a uma listagem das palavras com seus respectivos índices de ocorrência no corpus, seguido por uma ponderação das palavras (*weighting scheme*) e por fim na poda do dicionário ponderado.

Seja $F \in M_{m,n}(\mathbb{R})$ a matriz de freqüência absoluta, ou contagem, onde m é o número de documentos processados e n é a dimensão léxica. Então f_{ij} é o número de vezes que o termo j aparece no i-ésimo documento, $1 \leq i \leq m$, $1 \leq j \leq n$ e $f_{ij} \in \mathrm{N}$ (naturais).

Definimos o vetor-linha $d_i = F_{i\bullet} \in \mathrm{N}^n$ é a representação vetorial do i-ésimo documento: $d_i = \begin{bmatrix} f_{i1} & f_{i2} & \cdots & f_{in} \end{bmatrix}$ e $D = \{d_i\}_{i=1}^{m}$ o conjunto de documentos no corpus. Da mesma forma, o vetor-coluna $f_j = F_{*j} \in \mathbb{N}^m$ representa as ocorrências da j-ésima lexia nos m documentos: $l_j = [l_{1j} \quad l_{2j} \quad \ldots \quad l_{mj}]^T$, onde T sobrescrito indica a transposição do vetor e $L = \{t_f\}_{j=1}^{n}$ o conjunto léxico, ou dicionário.

As transformações descritas nos tópicos seguintes, referenciadas na literatura de IR por métodos de pesagem de termos (*term weighting schemes*), são aplicadas sobre a matriz F. Seja W uma transformação selecionada, digamos, $W = TF$ (Term Frequency). Então, para uma W qualquer,

$$W: M_{m,n} \rightarrow M_{m,n}$$

De forma que, por conveniência de notação,

$$W_{i,j} = W(F_{i,j})$$

A freqüência de termos (TF) é o mais simples e corriqueiro dentre os índices, e é muito usado na computação de outros índices mais sofisticados. Contudo, Sebastiani (1999) mostrou que mesmo o TF se provou ser muito eficiente. Abaixo temos a definição de diversos índices de ocorrência vistos na literatura de IR.

Antes da seleção de atributos, as matrizes F e W "possuem" uma coluna para cada termo existente na base, ou seja, têm a mesma dimensão do dicionário universal. Na realidade, se N é a dimensão léxica total, é computacionalmente exaustivo calcular $F, W \in \mathbb{M}_{m,N}$. O capítulo 5.2.2 sobre Pré-Processamento mostra uma implementação por linhas (documentos) mais eficiente.

Na Representação de documentos (*Document Representation)* (seção 2.6.3), a matriz de transformação $W \in \mathbb{M}_{m,n}$ é a saída final do pré-processamento e entrada para as ferramentas de indução (interpolação, extrapolação, etc.), com dimensão léxica reduzida a n.

Uma explicação detalhada sobre cada método de pesagem usado nesse trabalho é dada a seguir.

2.6.2.1 B: Boolean Weighting

De acordo com Salton e Buckley (1988), para o vetor d_i, o método de pesagem booleana considera $B_{i,j} = \mathrm{I}_{\{f_{i,j}>0\}}$ para um dado documento j, onde I é a função indicadora aplicada sobre o conjunto dado. Imediatamente derivamos a matriz ponderada $W = B$ (ou ainda $W_{i,j} = B_{i,j}$).

2.6.2.2 TF: Term Frequency

Com as definições acima, TF pode então ser definida como o vetor normalizado, com freqüência de um termo ponderada em relação aos demais termos do mesmo documento:

$$TF_{i,j} = \frac{f_{i,j}}{\sum_{j=1}^{n} f_{i,j}}$$

O uso de TF assume que termos importantes ocorrem com maior freqüência que os menos importantes (BRÜCHER, KNOLMAYER e MITTERMAYER, 2002).

2.6.2.3 TF-IDF: Produto TF x IDF

A freqüência de documento (DF) de um termo é o número de documentos em que ele ocorre, ou seja:

$$DF_j = \sum_{i=1}^{m} \mathrm{I}_{\{f_{i,j}>0\}}$$

Note-se a equivalência da notação de conjuntos usada por outros autores:

$$\sum_{i=1}^{m} \mathrm{I}_{\{f_{i,j}>0\}} = \left|\left\{d_i \in D : t_j \in d_i\right\}\right|$$

DF é um método puramente empírico i.e. não tem nenhum embasamento teórico (YANG e PEDERSEN, 1997). Os termos são podados a partir de um limiar inferior.

Uma vez definida a matriz *F* de freqüências de termos, podemos definir a freqüência inversa (IDF) do documento *i* como(YANG e CHUTE, 1994):

$$IDF_j = \log\left(\frac{m}{1+DF_j}\right)$$

A constante 1 é colocada simplesmente para evitar a divisão por 0, no caso do conjunto denominador ser vazio. Ao aplicar o *IDF* supõe-se que as palavras mais raras têm maior poder explanatório.

O índice $TF \times IDF_{i,j} = TF_{i,j} IDF_j = W_{i,j}$ é muito popular na literatura de IR. A qualidade dele está em valorizar palavras não excessivamente freqüentes (BRÜCHER, KNOLMAYER e MITTERMAYER, 2002).

2.6.2.4 IG: Information Gain

O ganho de informação mede o número de bits de informação obtido através da presença ou ausência de um termo num documento(FELDMAN e SANGER, 2007). Seja $\{c_k\}_{k=1}^{L}$ o conjunto de categorias no espaço-alvo, com *L* igual ao número de categorias distintas. O ganho de informação do termo *j* é definido como:

$$IG_j = -\sum_{k=1}^{L} \mathrm{P}(c_k)\log \mathrm{P}(c_k) + \mathrm{P}(j)\sum_{k=1}^{L} \mathrm{P}(c_k \mid j)\log \mathrm{P}(c_k \mid j)$$

A estimação da probabilidade tem complexidade de tempo de $O(m)$ e de espaço de $O(mn)$. O cálculo da entropia tem complexidade de tempo de $O(nk)$ (YANG e PEDERSEN, 1997). Assim, $W_{i,j} = F_{i,j} .* \left(1_{m,1} * IG_j\right)$, o operador .* corresponde à multiplicação por elementos pareados das matrizes.

2.6.2.5 MI: Mutual Information

Considere a seguinte tabela de contingência com as ocorrências do termo t_j e a categoria c_k (SINGH, MURTHY e GONSALVES):

Tabela 2.2 Tabela de Contingência para Termos e Categorias

Termo \ Categoria	k	$\bar{k}$
j	A	B
$\bar{j}$	C	D

O valor *A* é o número de vezes que o termo t_j e a categoria c_k co-ocorrem, *B* é o número de vezes que t_j ocorre sem c_k e C é o número de vezes que c_k ocorre sem t_j. $MI_{j,c}$ tem o valor natural de zero se t_j e c_k são independentes.

Para o termo *j* e a categoria *k*, o critério de Informação Mútua *MI* e sua estimativa são:

$$MI_{j,k} = \log\left(\frac{\mathrm{P}(j \cap k)}{\mathrm{P}(j)\mathrm{P}(k)}\right) \approx \log\left(\frac{mA}{(A+C)(A+B)}\right)$$

Para medir a qualidade de um termo na seleção de atributos globais, é possível combinar a pontuação específica por categoria de dois modos alternados:

$$MI_j^{med} = \sum_{k=1}^{L} P(c_k)\, MI_{j,k}$$

$$MI_j^{max} = argmax_{k=1}^{L}\{MI_{j,k}\}$$

Finalmente, $W_{i,j} = F_{i,j} .* \left(1_{m,1} * MI_j^{max}\right)$ ou $W_{i,j} = F_{i,j} .* \left(1_{m,1} * MI_j^{med}\right)$. Neste trabalho usaremos MI^{max}.

A fraqueza da informação mútua é que a pontuação é muito influenciada pelas probabilidades marginais dos termos. Para termos com probabilidade condicional $\mathrm{P}(j \mid k)$, termos raros terão maior pontuação que termos comuns (Yang, Pedersen).

2.6.2.6 Poda de Palavras

A poda de palavras elimina os termos com menor poder de discriminação, ou seja, aquelas com menor probabilidade de separar as categorias. Assume-se que sejam aqueles com freqüência muito baixa, ou após a ponderação dos vetores com os métodos descritos acima, com pesos reduzidos.

Em geral, 5% a 10% das palavras com freqüência superior são conservadas sem perda de desempenho do classificador, ou até mesmo ganho no poder de classificação em função da eliminação de ruído no conjunto de treinamento (YANG e PEDERSEN, 1997). A Figura 2.3 extraída do trabalho dos autores ilustra o potencial do corte de dicionário. As três primeiras curvas representam Information Gain (IG), Document Frequency (DF) e Term Strength (TS).

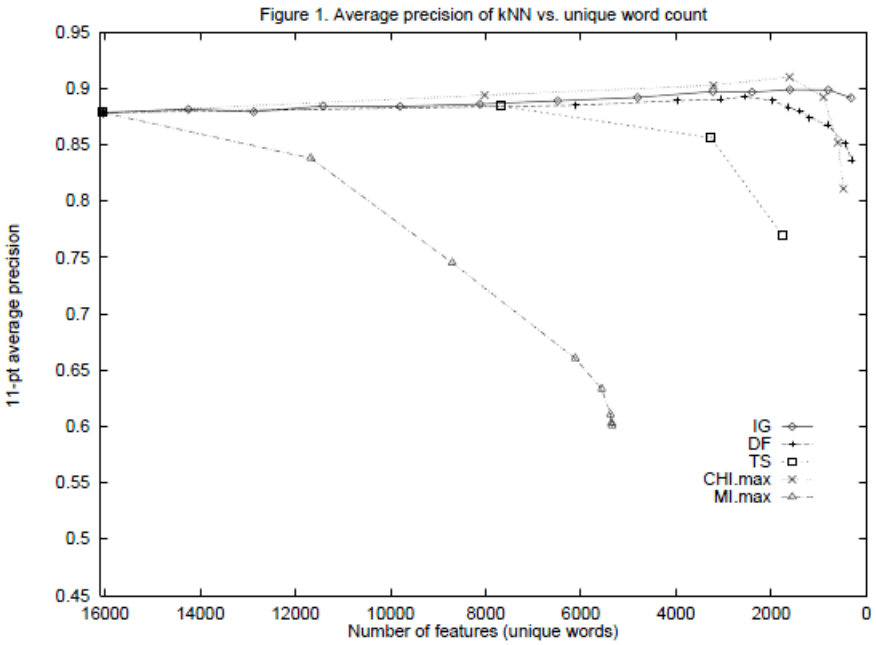

Figura 2.3 Ilustração da Poda de Dicionário (YANG e PEDERSEN, 1997)

A figura mostra o cruzamento do desempenho de precisão (entre 0 e 1) para tamanhos diferentes da dimensão léxica, isto é, para o número de palavras-chave usado. É visível que alguns métodos de pesagem tem desempenho consistentemente superior ao de outros, comparando-se suas curvas de desempenho. O objetivo nesta etapa é reduzir ao máximo a dimensão léxica com pouco comprometimento do desempenho. Neste estudo comparativo, uma redução de 98% da dimensão léxica foi alcançada com aplicação o método de Information Gain (IG).

2.6.3 Representação de documentos (Document Representation)

O objetivo da fase de Representação de Documentos é montar uma matriz de freqüência $F \in \mathbb{M}_{m,n}(\mathbb{R})$ contendo a contagem de termos selecionados por documento. A matriz equivale àquela descrita na seção 2.6.2, porém com uma dimensão léxica (n) reduzida.

Além disso, a representação de documentos atribui a cada documento d_i uma saída y_i desejada. A maioria das pesquisas atribui um rótulo da classe a qual d_i pertence, pareando cada observação à categoria desejada. No caso de previsão em FIR, é usual dividir as categorias em 2 rótulos, um para períodos de “Alta” e outro para “Baixa”. Ou adicionar um terceiro rótulo intermediário para períodos quando o ativo está “Estável”. Porém Schumaker e Chen (2009) parearam os documentos diretamente com o retorno da ação para aplicar regressão sobre os dados.

2.7 Indução ou Processamento (Induction)

O processamento dos dados é a aplicação da matriz F em uma máquina de associação de padrões, como regressão linear (GOODHART, HALL, *et al.*, 1993), k-nearest neighboors (WÜTHRICH, PERMUNETILLEKE, *et al.*, 1998), sistema baseado em regras (PERAMUNETILLEKE e WONG, 2000), support vector machine (FUNG, YU e LAM, 2002) e outros.

CAPÍTULO 3 NOTÍCIAS FINANCEIRAS

Foi visto anteriormente que FIR, como área de pesquisa, pode ser considerada um desmembramento de IR, uma vez que herdou metodologias e ferramentas desta. No entanto os objetivos de ambas são distintos: IR procura aperfeiçoar ferramentas de busca e indexação de textos e oferece uma base para análise de conteúdo, enquanto FIR tenta estabelecer uma relação entre os conteúdos textuais e o mercado financeiro.

Em função disso é que se nota a dualidade nos módulos de seleção de fontes e pré-processamento: são necessários 2 bancos de dados, um com notícias e outro com cotações (ver Figura 2.2). Como conseqüência, FIR possui um módulo extra para pareamento das notícias com as séries temporais, intermediário entre o pré-processamento dos textos e a indução.

3.1 Tipos de Notícias

A variedade de notícias que influenciam os mercados financeiros é estratificada por Johnson (2010) da seguinte forma:

- Global / Regional: Eventos políticos, guerra e terrorismo, desastres naturais, manifestações populares
- Macroeconômicos: Taxa de juros, renda, desemprego, expectativa de crescimento do PIB
- Corporativas: Fusões e aquisições, falência, mudança no quadro executivo, lançamento de produtos, relatórios trimestrais e anuais, anúncio de dividendos

O grupo de notícias que não afeta os mercados é ruído de informação. É certo que dentre os estratos descritos, este compõe o maior volume de informação documentada na internet. Por isso a seleção adequada da fonte de notícias e posterior filtragem das notícias capturadas é essencial em qualquer modelo de *News Trading*.

Uma quarta fonte de informações que vem sendo estudada mais recentemente são as *Message Boards*, ou grupos de debate da internet. São comunidades online que discutem notícias, rumores e perspectivas sobre o mercado de maneira informal.

Dos 3 grupos influentes, as notícias globais e regionais são as mais difíceis de interpretar (JOHNSON, 2010). Por outro lado, as notícias macroeconômicas são mais simples, por serem periódicas e de conteúdo minimalmente numérico: podem ser resumidas a um ou poucos números que generalizam aproximadamente a totalidade da informação passada.

As notícias corporativas mesclam propriedades de ambos os grupos: podem ser periódicas e minimais, como relatórios trimestrais, anuais e dividendos, mas são em maioria aperiódicas e plenamente textuais.

3.2 Notícias Corporativas

A tese de Thomas (2003) divide as notícias corporativas em 11 classes, contendo seus respectivos tipos de notícias:

- Atividade de Analistas: Cobertura de analista iniciada, Upgrade de analista, Downgrade de analista, Upgrade de rating, Downgrade de rating, Análise de ação, Análise de setor, *Conference Call*
- Controle Corporativo: Fusão, Aquisição, *IPO*/*Spin-Off*, *Tracking Stock*, Investimento minoritário, Reunião de acionistas, Assembléias gerais, Desinvestimento, Falência
- Assuntos Legais e Regulatórios: Ação judicial, Ação regulatória, Arquivamento anual na SEC (10-K), Arquivamento trimestral na SEC (10-Q), Arquivamentos intra-trimestrais na SEC (8-K)

- Resultados: Aviso de Resultados (*Warning*), Publicação de resultados (*Earnings Report*), Perspectivas de ganhos (*Earnings Outlook*), Revisão dos lucros (*Earnings Restatement*), Dedução dos resultados (*Earnings Charge*)
- Atividade do Mercado de Capitais: Novo financiamento, Recompra de ações, Repagamento de dívida, *Stock Split*, Opções
- Atividade Operacional (Produção/Vendas): Aquisição de produto, Campanha de vendas, Vendas de produtos, *Recall* por defeitos, Pesquisa de mercado, Licenciamento, Corte de preço, Aumento de preço
- Atividade da Ação: Novidades sobre o desempenho das ações, Atividade setorial, Atividade de mercado, Atividade das commodities
- Relações Públicas: Marketing, Comentários do CEO, Caridade, Apresentação/Discurso públicos, Prêmios
- Relações corporativas: Relações de negócio, Relação com fornecedores, Parceria com grupo industrial, padronizações (*Standards*)
- Assuntos de Gestão Interna: estórias sobre o funcionamento interno de uma empresa e seus negócios, tais como: corte na folha de pagamentos; mudança na direção; reestruturação; compensação de executivos; assuntos trabalhistas; produção
- Miscelânea: Insider trading, Desastres naturais, Convulsões políticas (e outros fatos que afetem as perspectivas da empresa), Dividendos, Analise econômica, Atividades de fundos, Analise de fundos, Ajuste de índice, Venda de ativos, Certificação, Hackers, Vírus, etc.

O autor classificou um grande conjunto de notícias manualmente, de forma que as classes apresentadas acima sintetizam com eficácia os tipos de notícias encontrados numa base de dados financeira. O interesse nessa classificação está apenas em observar os tipos de notícias que costumam ocorrer no *corpus*.

3.2.1 Exemplo descritivo

Como ilustração do impacto de comunicados à imprensa, tomemos o evento de divulgação dos resultados do primeiro trimestre fiscal de 2012 da Nike (Nike, Inc. – NKE) com ações negociadas na NYSE. O comunicado foi publicado através da Business Wire, como mostra a manchete abaixo encontrada no site da agência.

Figura 3.1 Exemplo de divulgação dos resultados trimestrais –Nike Inc.

De acordo com o relatório de divulgação de resultados da empresa publicados no site da Bloomberg | Business Week, a estimativa consensual pré-divulgação de 18 analistas sobre o EPS era de $1.22 dólares norte-americanos, variando num intervalo de $1.13 a $1.38. O resultado divulgado no dia 22 de Setembro de 2011, às 16:15 (horário local UTC/GMT-5h + DST[1h]) de $1.36 foi 11,48% acima da média consensual. É importante notar que o resultado chegou ao público logo após o encerramento do pregão da NYSE, que ocorre às 16:00 do horário local.

No intervalo do momento de abertura do pregão do dia seguinte (23/set/2011, 09:30) até 30 minutos subsequentes, o valor da ação NKE subiu 6% em relação ao fechamento do dia 22, como o gráfico abaixo mostra (retirado de http://www.freestockcharts.com/ ©2010 Worden Brothers Inc.).

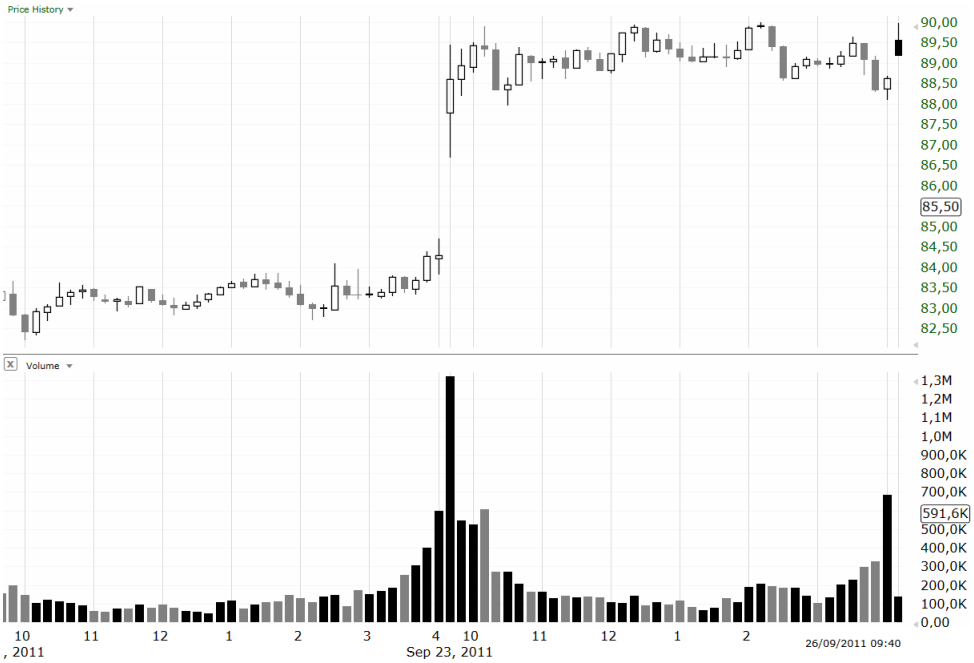

Figura 3.2 Preço e volume de NKE antes e após divulgação de resultados

A aparente antecipação dos resultados imediatamente antes da divulgação já foi analisada em diversos artigos (JAMES M. PATELL, 1979), (JENNINGS e STARKS, 1986). Sobre a antecipação, é possível especular sobre a possibilidade do mercado antever os reultados com base nas informações públicas disponíveis até minutos antes da divulgação, ou a respeito de um potencial *Insider Information Trading*.

3.2.2 Política de Divulgação

No Brasil, a Comissão de Valores Mobiliários (CVM) tem o mesmo papel de regular o mercado financeiro da SEC (*Securities and Exchange Commission*) nos Estados Unidos. De acordo com o Manual da Política de Divulgação (BOVESPA, 2008), seção V - Política de Divulgação de Ato ou Fato Relevante, artigo 3, o objetivo da divulgação de ato ou fato relevante:

> *Art. 3: (...) assegurar aos investidores a disponibilidade, em tempo hábil, de forma eficiente e razoável, das informações necessárias para as suas decisões de investimento, assegurando a melhor simetria possível na disseminação das informações. Desta forma, impede-se o uso indevido de informações privilegiadas no mercado de valores mobiliários pelas pessoas que a elas tenham acesso, em proveito próprio ou de terceiros, em detrimento dos investidores em geral, do mercado e da própria Companhia.*

O acesso, a restrição e a divulgação de informações internas das companhias de capital aberto são rigorosamente controlados para promover a eficiência do mercado secundário de títulos. O controle da divulgação é definido no artigo 7. Quando Informar e Divulgar – Prazos:

> *Sempre que possível, a divulgação de Ato ou Fato Relevante deverá ocorrer antes do início ou após o encerramento dos negócios nas Bolsas de Valores, sendo que, em caso de incompatibilidade de horários com outros mercados, prevalecerá o horário de funcionamento do mercado brasileiro.*

Um vazamento antecipado de informação sigilosa da empresa (*insider information*) com potencial impacto sobre o preço da ação é considerado crime, e conhecido por *insider trading*.

De acordo com Mittermayer (2004), as companhias de capital aberto são requeridas pelo *Securities Exchange Act* de 1934 a garantir divulgação pública simultânea de informação privada relevante por causa da importância potencial de tais eventos para os investidores. Essas informações incluem divulgações de resultados, aquisição ou desinvestimento de negócios e saída do cargo pela diretoria. Quase todas as sociedades abertas nos Estados Unidos têm contrato de divulgação de *press releases* com empresas de comunicação para preencher os requisitos exigidos pela regulamentação. As duas maiores, PR Newswire e Business Wire dividem o mercado em quase meia fatia cada.

O mesmo autor argumenta que em função da política de divulgação, *press releases* são uma boa fonte de informação, por serem o canal primário de divulgação e por tratarem de informações relevantes. Notícias que não se encaixam nessas características podem ser publicadas em outros canais comuns.

3.3 Trabalhos Anteriores em FIR

Nesta seção faremos uma revisão da coletânea de todos os trabalhos encontrados diretamente relacionados com FIR, do primeiro, de 1997, até 2009. No APÊNDICE B (página 148) se encontra a tabela de consulta rápida para os trabalhos discutidos a seguir. Nela se apresenta um resumo da literatura completa, com descrição da metodologia, resultados e comentários. As primeiras colunas da tabela contêm metadados sobre os trabalhos (ano, autor, publicação, local) e as demais decompõem algumas características principais da metodologia.

A primeira aplicação das técnicas de IR em notícias financeiras foi conduzida por Peramunetilleke em sua dissertação de mestrado em Ciência da Computação pela Universidade de Hong Kong em 97 (PERAMUNETILLEKE, 1997). Nesse trabalho, Peramunetilleke realiza a primeira previsão de movimentos de um ativo financeiro com base apenas no texto das manchetes de notícias do mercado monetário. Os ativos estudados foram as taxas de oferta de compra (*bid*) das taxas de câmbio Marco alemão / Dólar americano (DEM/USD) e Iene japonês / Dólar americano (JPY/USD).

A metodologia emprestou técnicas de pré-processamento de textos desenvolvidas em IR. Foram usados 3 métodos de pesagem (B, TFxIDF e TFxCDF) além de *principal component analysis* (PCA) para seleção de n-uplas de palavras. Para previsão foi desenvolvido um algoritmo simples de geração iterativa de conjuntos de regras probabilísticas baseadas nas n-uplas, que minimizem o erro de estimação.

Como resultado do trabalho, o autor mostrou que é possível extrair informação quantificável dos textos. Além disso, os métodos de pesagem tradicionais em IR são mais simples, computacionalmente econômico e provêem uma acurácia superior ao PCA.

Ao trabalho iniciado por Permunetilleke em 97 foi dada continuidade por seu orientador Wüthrich, em conjunto com Leung, Cho, Zhang e Lam no ano seguinte (WÜTHRICH, PERMUNETILLEKE, *et al.*, 1998). Nele, os autores programaram um modelo de mineração de informações veiculadas na web – notícias do Wall Street Journal, Financial Times, CNN, Bloomberg, Reuters e Dow Jones – com pré-processamento léxico para previsão diária de cinco índices: DJI, Nikkei 225, FTSE 100, HSI e STI.

Neste trabalho foi introduzido o método "b*ag of words*", que incorpora o conhecimento prévio de especialistas para a seleção de um conjunto estático de palavras-chave que serão usados para a etapa de mineração de textos.

O objetivo consistia em classificar *ex-ante* o movimento dos índices entre ‘up’, ‘steady’ e ‘down’, sendo que o intermediário equivalia ao intervalo [-0,5%; 0,5%]. Para tal, foi utilizado uma dimensão léxica de 423 termos-chave e 3 classificadores para classificação: sistema baseado em regras probabilísticas, kNN com 9 centróides e MLP retropropagada de saída binária com 211 nós ocultos. Os resultados de previsão foram positivos e favoráveis à primeira técnica.

Ainda em 98 na Univeridade de Michigan, Wysocki estudou a relação entre o volume de postagens em fóruns de discussão da internet sobre ações (WYSOCKI, 1998). Descobriu, com base puramente em estatística descritiva, que a quantidade de mensagens postadas tem relação positiva direta com a exposição de uma ação (volume de operação, capitalização da empresa no mercado e cobertura por analistas) e com seu risco (resultados extremos publicados no passado, e volatilidade, véspera de anúncio de resultados).

Ainda mais interessante, o autor achou uma autocorrelação serial no volume de postagens: 1% no aumento de postagens de um dia leva a 0.23% de aumento do dia seguinte.

Da mesma forma, um aumento de 10x no volume de postagens no overnight leva a um aumento médio de 15.6% do volume de trade, e de 0.7% de retorno anormal (retorno da ação menos o retorno de mercado).

Em 1999 Cho, Wüthrich e Zhang voltam a escrever sobre o trabalho iniciado no ano anterior, dessa vez focando sobre a etapa de pré-processamento dos textos, que é fundamental na extração do conteúdo semântico das notícias. Estudam quatro esquemas de medida dos vetores de entrada baseados em *Term Frequency* (TF, ou contagem de termos-chave). Na literatura de IR, termos diferem de palavras-chave por serem "n-uplas" (ou tuplas) de palavras (e.g. 'bond rise' ou 'share drop'). Os autores comparam os esquemas pela previsão diária do HSI (Hang-Seng Index), novamente utilizando as classes up, steady e down.

O trabalho conclui (por meio de análise qualitativa dos resultados) que o esquema de '*vector weighting with cluster relevance and discrimination*' possui melhor desempenho que os demais métodos e relativamente melhor que '*simple weighting*'. Não faz parte do escopo deste trabalho revisar quais são os esquemas adotados, pois todos são aplicáveis exclusivamente a problemas de classificação, e aqui estaremos trabalhando com um conjunto-imagem contínuo. No entanto, alguns resultados são importantes. Primeiro, a probabilidade de que a acurácia de previsão randômica seja maior ou igual à do pior esquema de medida foi de apenas 0,18%. Segundo, ao testar a remoção de termos do léxico, observou-se que o desempenho reduz "sensivelmente" apenas a partir de um léxico de 392 termos, e com a retirada das palavras "bonds", "stock", "currency" ou "inflation".

Lavrenko et.al. (2000) adotaram uma rota alternativa à evolução da pesquisa de até então, desenvolvendo um sistema que faz recomendação da leitura de notícias – aquelas publicações com um conteúdo lingüístico mais alinhado a uma determinada classe de artigos são selecionadas – e mostram que a recomendação dos textos pelo programa pode ser rentável.

Foi utilizado um mecanismo para quebra da série temporal em tendências, chamado *Piecewise Linear Regression* ou conhecido também por *Piecewise Segmentation*. O algoritmo consiste quebrar uma série temporal em pedaços recursivamente menores, nos pontos em que a quebra maximiza uma determinada medida de ajuste ou até atingir um ponto de parada. Neste trabalho o teste-t foi usado como ponto de parada. Num segundo passo, os autores mediram o comprimento, inclinação, intercepto e R^2 de cada segmento e nomearam-nas subjetivamente de acordo com tais características. Os rótulos adotados foram 'SURGE',

‘SLIGHT+’, ‘NO RECOMENDATION’, ‘SLIGHT–‘ e ‘PLUNGES’. Como método de inferência sobre os textos minerados, foi aplicado uma modelagem de linguagem bayesiana sobre todas as notícias publicadas até 5 horas antes do período previsto.

Os resultados auferidos indicam um resultado muito superior da metodologia desenvolvida em relação à tradicional (espaço vetorial com TFxIDF) e à classificação aleatória, o que sugere que o modelo tradicional não é capaz de capturar as relações entre as notícias e as tendências no preço. Com relação à simulação de trading, aplicando um conjunto de regras simples de *shorting* a sinais negativos e vice-versa, o sistema obteve um retorno médio de 0.23% por operação, significativo ao nível de 1% de confiança. O retorno médio por operação foi decrescente à medida que se aumenta o intervalo de entrada dos dados no sistema: 0.03%, 0.02% e 0.16% para 1, 5 e 10 horas respectivamente.

Adicionalmente, este estudo comparou o uso de treinamento baseado no conjunto completo de notícias com o uso de notícias específicas para uma empresa e verificou que notícias específicas levam a uma previsão mais acurada sobre a direção do movimento de preço (Schumaker e Hsinchun, 2009 apud Lavrenko et.al., 2000)

O estudo de Thomas e Sycara (2000) foi importante para fechar o elo ligando o estudo de Wyocki (98) à linha de mineração de textos iniciada por Peramunetillek. Nele, os autores desenvolveram 2 metodologias para estudar fóruns de discussão online: a primeira de previsão por classificação com máxima entropia usando somente todo o conteúdo textual das postagens, e a segunda um algoritmo genético para aprender regras simples de *trading* baseadas somente em dados numéricos (volume de trading, número de mensagens postadas e número de palavras postadas).

A primeira abordagem teve um resultado de trading simulado (+6.91%) ligeiramente melhor que a segunda (+5.94%) ao longo de 200 dias de teste. Contudo, a mescla de ambas produziu um resultado superior de cerca de 20%. No entanto, é importante notar que o uso feito de 2 classes para classificação não é capaz de distinguir conteúdo textual ruidoso ou mesmo espúrio. Além disso, a simulação rodou sobre apenas 52 observações de treinamento com cotações de fechamento, o que pode representar uma janela curta de padrões complexos para serem aprendidos, ou seja, problema de alta dimensionalidade.

Tal como seus colegas, Peramunetilleke e Wong (2000) reaplicaram a metodologia de 97, usando mais uma vez o sistema de previsão baseada em regras com 3 classes. Foram

adotadas as taxas DEM/USD e JPY/USD como objeto-alvo de análise mais uma vez. Os resultados no campo foram pouco ampliados pela metodologia empregada ser muito semelhante ao trabalho anterior. Contudo a hipótese nula de previsão aleatória foi testada, com um nível descritivo de 0,4% para acurácia de 51% com 3 classes, amostra de 60 e aproximação da distribuição binomial pela normal.

Gidófalvi (2001) confirma que os movimentos de curto prazo no preço de ações podem ser previsto a partir de artigos de notícias financeiras, apesar do poder preditivo da ferramenta desenvolvida ser baixo. O autor argumenta que dentre os 20 min anteriores e posteriores ao lançamento do artigo, é possível prever o movimento de um ativo, o que contrariaria a EMH.

Seguindo a prescrição anterior, a classificação é feita baseada nos rótulo “up”, “down” e “unchanged”, dadas às notícias de acordo com o movimento do preço do ativo ao longo do período subseqüente, para fins de previsão. A ferramenta utilizada foi um *Naïve Bayes Classifier*. O autor modela o impacto das notícias a partir de uma janela delimitada por um deslocamento inferior e outro superior, ambos pendentes a partir do instante de publicação da notícia, com o objetivo de verificar a eficiência de mercado no que diz respeito à velocidade com que a notícia é absorvida pelo mercado.

Os resultados do teste Qui-Quadrado contra a hipótese nula, de que os resultados do experimento são iguais aos resultados de previsão aleatória, indicam um grande nível descritivo para deslocamentos dentre 30 minutos, à esquerda e à direita do instante de publicação. Sobre tal resultado, o autor faz referência a Elkan (1999) questionando a possibilidade de diferentes fontes de notícia publicarem a mesma informação em momentos espaçados, o que explicaria a gradual absorção da informação no intervalo dentro da moldura da EMH.

Fung, Yu e Lam (2002) detalham um novo procedimento para inferência sobre a série de cotações e notícias. Em primeiro lugar, desenvolvem um novo algoritmo de descoberta de tendências, ou *piecewise segmentation algorithm*, consistindo de uma segunda fase de agregação, oposta à primeira fase de quebra, o que segundo eles evita a sobre-segmentação da série. A segmentação é seguida pela rotulação dos segmentos. Nesta etapa, cada segmento é tratado como um objeto, e sobre eles um algoritmo de aglomeração é aplicado, unindo conglomerados de objetos de acordo com a medida de distância mínima média do grupo (GAD, *minimum group average distance*). O processo de aglomeração termina ao alcançar 3

aglomerados, rotulados como “Rise”, “Drop” e “Steady” conforme o coeficiente de inclinação médio do grupo de segmentos.

A terceira etapa, de alinhamento de notícias com tendências, também contém um novo algoritmo introduzido pelos autores nomeado aglomeração guiada (*guided clustering*), que resolve o problema de alinhamento de notícias espúrias com as tendências através de filtragem. É uma extensão do tradicional k-means incremental.

Os autores propõe o uso de dois novo esquemas de pesagem dos vetores, CDC (inter-cluster discrimination coefficient) e CSC (intra-cluster similarity coefficient). Comparar a relação com CDF, ou demais já usados. Support Vector Machine (“SVM”) é a ferramenta usada para inferência. Os resultados mostram que o sistema desenvolvido é rentável, porém os números não são divulgados, de forma que não é possível avaliar quanto questões como benchmark e cenário econômico poderiam anular os retornos positivos do modelo.

Seo, Giampapa e Sycara (2002) e Seo, Giampapa e Sycara (2004), desenvolveram um sistema que realiza a recomendação de notícias financeiras classificadas de acordo com cinco categorias: “good”, “good uncertain”, “neutral”, “bad uncertain” e “bad”, através de aprendizagem supervisionada sobre 1.239 artigos de notícia rotulados manualmente. Uma estrutura de representação léxica igual a das publicações anteriores é empregada, baseada na matriz de freqüência, com o índice TFxIDF normalizado, ou *bag of words*.

O método Domain Experts desenvolvido prediz a classe de um artigo financeiro através de um processo de votação de agentes especializados no status de uma empresa de interesse, com uso de frases co-localizadas. Uma frase co-localizada é definida como uma seqüência de palavras próximas, mas não necessariamente consecutivas. As frases co-localizadas foram selecionadas por tentativa e erro com medição do ganho de informação (não há maiores detalhes). O artigo apresenta poucos detalhes a respeito do mecanismo por trás dos agentes especializados, como numa abordagem aparentemente comercial.

A tese de doutoramento de Thomas (2003) expõe uma primeira seção sobre simulação de estratégias de trading baseadas em análise técnica automatizada. O conteúdo mais interessante desta seção está na abordagem da liquidez dada aos resultados das simulações. A liquidez traduz no problema da execução das ordens colocadas. Para aferição precisa da execução de ordens de uma estratégia automatizada, é necessário levar-se em conta a série de ofertas de compra e venda (*bid-ask*), além da série de preços. O custo computacional pode ser

muito elevado, ou ainda, a série pode não estar disponível. Contudo, o autor expõe uma análise dos retornos das diversas estratégias de acordo com os quintis de liquidez (em termos de volume negociado) das ações simuladas. Com isso foi possível verificar quais estratégias tinham melhor desempenho a que nível de liquidez, o que facilita a decisão de investimento.

Além do estudo sobre indicadores técnicos, o autor dedica um capítulo a mensagens postadas em fóruns de discussão online como feito por (WYSOCKI, 1998). Estatísticas descritivas são apresentadas usando um conjunto de postagens do Yahoo! Message Boards do início de 98 a agosto de 2001 que corroboram os achados anteriores.

O capítulo seguinte usa manchetes de notícias também publicadas no Yahoo!, e o autor tenta classificá-las entre 40 tipos de notícia definidos à mão pelo autor para os conjuntos de treino e teste. Os resultados da classificação foram positivos em termos de precisão, variando próximo de 90% para a maioria das classes de notícias, mas medianos para revocação, que variou bastante entre cada classe. Simulações de trading com indicadores técnicos sobre as notícias também não tiveram resultados extensivamente positivos, exceto pelas notícias relacionadas a divulgação de resultados da empresas.

Kroha e Baeza-Yates (2004) propõem utilizar a mineração de textos para classificação de segmentos de tendência temporal nos índices da DAX30 e da Nasdaq100. São delimitados 4 segmentos manualmente (Down2000, Down2002, Up1999 e Up2003) que são então convertidos em classes. As notícias são alinhadas com tais classes de acordo com o momento de publicação delas. Aplicando o Naïve-Bayes Classifier sobre o IDF dos documentos, os autores observaram 90% de acurácia, o que os leva a concluir que cada período de longo prazo de tempo pode ser identificado por um conjunto característico de ocorrência de palavras.

Mittermayer (2004) usou uma filtragem inicial das notícias que poderiam causar ruído ou redundância no sistema: notícias sem *ticker*, com mais de 1 *ticker*, que não fazem referência à bolsa onde a empresa está listada, que faz referência a outras bolsas que não das cotações coletadas, e publicadas fora do intervalo das 9:30 as 15:00 (da abertura a uma hora antes do fechamento) para ações da NYSE e 8:00 as 17:00 para ações da NASDAQ. Essas restrições limitaram a base de notícias a 6.602.

Para rotular “Good News”, o autor usou uma definição própria: notícias que seguem de um aumento em 1% médio nos 60 minutos seguintes e que em algum ponto desse intervalo

levou a um pico de 3%. O mesmo aplica às “Bad News”, e “No Movers” é o conjunto complementar de ambos. Com isso o sistema foi capaz de gerar automaticamente seu próprio dicionário, eliminando a necessidade de um especialista para a geração do *bag of words* estático.

Usando um pacote de *Support Vector Regression*, o sistema teve uma revocação média de 58% entre as três classes, mas precisão média inferior a 33%. Em simulação de operação, o retorno foi de cerca de 0,20% por *trade* contra 0,07% médio de um *trader* aleatório.

Em Mittermayer e Knolmayer (2006) os autores apresentam uma revisão teórica dos resultados obtidos pelos protótipos testados até então na bibliografia de FIR. Os autores observam que apenas 3 dos 8 protótipos desenvolvidos até então em FIR usam dados intradiários, e que isso não representa ganhos em desempenho.

Duas métricas de desempenho de classificação, ambas provindas de IR, foram adotadas: precisão (*precision*) e revocação (*recall*). Apesar de o autor relatar taxas de precisão e revogação baixas para as diferentes categorias, um experimento de aplicação do sistema – isento de custos transacionais – mostrou que seria possível obter retornos consistentemente positivos em cerca de 2.600 operações de compra e venda baseada em regras simples, com retorno médio de 0.11% por operação.

No mesmo ano apresentaram um novo protótipo e seus resultados (MITTERMAYER e KNOLMAYER, 2006). O protótipo empregava uma metodologia semelhante à do estudo de Mittermayer de 2004, com a mesma base de notícias e com pré-filtragem. No entanto foram selecionadas apenas notícias dentro das categorias com maior potencial de impactar o mercado: “Dividendos”, “Projeção de Resultados”, “Acordos de Financiamento”, “Assuntos Legais”, “Acordos de Licenciamento”, “Ofertas” e “Relatório de Vendas”. Pela primeira vez em FIR, foram realizados diversos testes cruzados sobre os parâmetros da metodologia: pesagens para seleção de atributos, tamanho do conjunto de atributo (dimensão léxica), pesagens para representação dos documentos e por fim, diferentes classificadores.

Para seleção de documentos foram comparados CTF, CTFxIDF, OR, CS e MI. Os resultados apontaram para o uso dos 3 primeiros, com 82 a 83% de acurácia direcional. Com relação ao tamanho do conjunto de atributos, foi medida a razão entre o número de atributos e o número de documentos. Entre 5 a 25% foi observada uma mudança quase nula no desempenho. Para o método de pesagem na representação, a acurácia foi quase igual entre

WDFxIDF, B, WDF e IDF, mas WDFxIDF foi escolhido como melhor apenas em função do resultado financeiro de *trading* simulado. Enfim, para inferência, SVM teve um desempenho um pouco superior (cerca de 82%) ao Rocchio (80%) e ao kNN (79%) para acurácia.

Schumaker e Hsinchun (2009) propõem o uso de previsões para setores industriais, em contraponto com os estudos anteriores que testaram notícias universais ou específicas para o objeto-alvo. As empresas foram rotuladas com base no sistema GICS (Global Industry Classification Standard) do banco Morgan Stanley, que possui 4 níveis hierárquicos: setor, grupo, indústria e sub-indústria.

Além disso, foi aplicada regressão com SVR sobre a representação vetorial das notícias. O valor futuro das ações foi estimado, com isso diferentes métricas de desempenho, das comumente usadas em IR, foram adotadas (e.g. precisão e revogação). Os resultados apresentaram melhor acurácia direcional (71,18%) e maior retorno simulado (8,50%) para a divisão de notícias por setor, porém o treinamento com notícias universais teve maior capacidade de acertar o preço futuro (MSPE = 0,0443). Todos os resultados se encontraram dentro do nível de confiança de 5%.

3.4 Hipótese do Mercado Eficiente

A teoria da eficiência de mercado assume que o preço dos ativos, a qualquer momento, "reflete completamente" toda a informação disponível (FAMA, 1970). Num exemplo intuitivo (adaptado de BODIE, KANE e MARCUS, 2000), se é possível prever hoje – com base nas informações disponíveis até o momento – que o preço de uma determinada ação subirá, digamos, 3% amanhã, então os investidores correrão a comprar a ação o quanto antes para usufruir do movimento de alta, o que levará o preço a subir 3% instantaneamente. Desta forma, a previsão não ocorre, pois o preço já reflete de todas as formas possíveis a informação disponível até o momento. Por outro lado, se a previsão ocorresse, ela não faria parte das informações de hoje.

Esse raciocínio simplifica a hipótese mais ortodoxa da eficiência de mercado. Contudo a literatura aponta 3 principais tipos de formas da eficiência, que se distinguem pela definição de "toda a informação disponível". Em primeiro lugar, a forma fraca, na qual toda informação disponível é somente histórica, a forma semi-forte, que além dos preços históricos incorpora

toda a informação publicamente disponível, e em último lugar a forma forte, que abrange também as informações privadas da empresa. Perceba que para qualquer forma, informações históricas são preditoras de valores futuros apenas se a hipótese for falsa.

Se assumirmos que toda a informação disponível implica que as mudanças sucessivas no preço são independentes e adicionarmos a hipótese de que as mudanças são identicamente distribuídas, então o modelo resultante seria de passeio aleatório.

O mercado eficiente aloca recursos (i.e. compra ações) em prol de seus usos mais produtivos (i.e. de empresas mais rentáveis) como resultado da competição entre investidores maximizadores de riqueza própria. Além do mais, as bolsas de valores criam mercados eficientes, que disponibilizam opções aos investidores de alocaram seus recursos. Como os investidores são maximizadores de riqueza, eles colocarão seus recursos à disposição de empresas que eles consideram terem maior produtividade, ou em outras palavras, que remunerem melhor seu capital investido. Somado a isso, pelo fato de todos terem sempre o mesmo acesso à informação, suas decisões de investimento convergirão ao preço de títulos que reflete o que se espera que seja seu preço “real”.

Um suposto mercado “perfeito” possui muitos investidores racionais que buscam maximizar retorno e minimizar o risco, cada um com as mesmas informações e expectativas com relação aos títulos; e nenhuma restrição sobre investimentos, impostos ou custos transacionais. Em suma, um mercado perfeito agrega todos os pressupostos que o permitam ajustar os preços continuamente.

Em suma, a eficiência dos mercados é um tópico importante em economia, pois se eles funcionam ineficientemente, então o capital está sendo literalmente mal alocado, gerando perdas na economia (THOMAS, 2003).

CAPÍTULO 4 PROCESSAMENTO E INFERÊNCIA

Este capítulo apresenta de forma introdutória o Perceptron Multicamadas (MLP), que é o modelo de rede neural utilizado ao longo do trabalho. É feita uma revisão sobre os dois métodos de minimização dos erros de estimação mais comumente usados com MLP: a descida do gradiente e o método de Newton. A seguir é dada uma amostra do algoritmo de Retropropagação (BP) para aprendizagem da rede neural e por fim o Gradiente Conjugado Escalado (SCG) que será usado nessa pesquisa é apresentado.

4.1 Perceptron Multicamadas

O perceptron multicamadas (MLP) é um modelo arquitetural precursor das redes neurais, e certamente o mais comum na literatura. O trabalho original de (RUMELHART, HINTON e WILLIAMS, 1986) descrevia o perceptron semelhante ao desenvolvido por (ROSENBLATT, 1958), mas resolvia um problema que há décadas a área de Inteligência Artificial procurava resolver: a separação de padrões não-lineares. O caminho já era conhecido: (MCCULLOCH e PITTS, 1943) já haviam provado algebricamente que qualquer padrão lógico, por mais complexo que fosse, poderia ser representado (leia-se resolvido) por uma determinada árvore de operações binárias. Logo já se sabia que por meio do encadeamento de diversos perceptrons lógicos de Rosenblatt (ver Figura 4.1), poder-se-ia resolver uma ampla classe de problemas. Contudo, a dificuldade estava em traduzir a aprendizagem discreta, univariada e unária (de apenas 1 neurônio) proposta por Rosenblatt numa apredizagem contínua, multivariada e n-ária para aplicação genérica(HAYKIN, 2001).

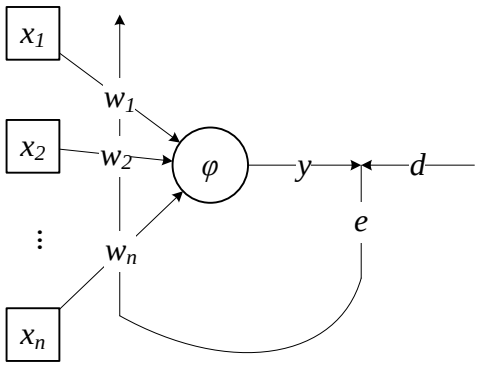

Figura 4.1 Perceptron de Rosenblatt

Rumelhart, Hinton e Williams propuseram o algoritmo de retro-propagação dos sinais de erro de aproximação através de uma rede com camadas de neurônios com funções de transferência contínuas, o que permitiu atribuir correções de pesos sinápticos simplesmente por meio do inverso do gradiente do erro retro-propagado. Essas correções podem ser traduzidas como a direção de maior minimização do erro de estimação que cada componente da rede neural pode realizar. Cronologicamente, o MLP não foi a primeira rede neural proposta. Um ano antes (ACKLEY, HINTON e SEJNOWSKI, 1985) publicaram a Máquina de Boltzmann. No entanto a MLP se popularizou mais, possivelmente pelo bom desempenho aliado à simplicidade do modelo.

4.1.1 Descida do Gradiente

É um algoritmo de otimização linear, que usa o vetor oposto ao gradiente, logo, a direção com a tangente de maior decrescimento da função, para buscar um ponto de minimização da função de custo.

Este algoritmo apresenta duas versões: uma estocástica e outra por lote (batch). A descida estocástica considera o gradiente referente a apenas um ponto do espaço amostral por iteração, enquanto seu primo por lote usa a soma dos vetores gradientes de todos as observações.

A descida estocástica leva esse nome por apresentar um comportamento ziguezagueante da solução do sistema (o vetor de parâmetros), o que ocorre em função da distribuição – aleatória – dos vetores de entrada. Apesar de perder o nível de paralelismo

sobre a apresentação das observações, em geral a descida estocástica converge mais rapidamente que o modo por lote.

Seja $\varepsilon(w(n))$ o custo de aproximação arbitrária, contínua e diferenciável em $\Re$, em função do vetor de parâmetros $w(n) \in \Re^m$ estimados da iteração *n*. Então o gradiente do custo é dado por $\nabla\varepsilon(w) = \left[\frac{\partial\varepsilon}{\partial w_1}, \frac{\partial\varepsilon}{\partial w_2}, \ldots, \frac{\partial\varepsilon}{\partial w_n}\right]$, com ∇ o operador gradiente por *w* apontando a direção de maior crescimento do custo.

O algoritmo de decida do gradiente é definido por:

$$w(n+1) = w(n) - \eta\nabla_w\varepsilon(n) \tag{1}$$

Que corresponde a um passo na direção oposta do gradiente. Observe que $\Delta w = w(n+1) - w(n)$. A constante η é uma constante suficientemente pequena de amortecimento da correção, introduzida no algoritmo por motivo empírico: ela regula o trade-off entre uma maior velocidade de correção (com valores maiores) e uma maior precisão de convergência.

A prova de convergência do algoritmo é simplesmente dada pela adaptação direta do polinômio de Taylor de primeira ordem:

$$\varepsilon(w(n+1)) = \varepsilon(w(n)) + \nabla\varepsilon(w(n))\Delta w(n) \tag{2}$$

Queremos que $\Delta\varepsilon(w(n))$ seja menor ou igual a zero, ou seja, a cada nova iteração o custo do erro de aproximação diminua. Passando $\varepsilon(w(n))$ para a esquerda, e da definição do algoritmo, temos

$$\Delta\varepsilon(w(n)) = \nabla\varepsilon(w(n))[-\eta\nabla\varepsilon(w(n))] = -\eta\|\nabla\varepsilon(w(n))\| \tag{3}$$

Como a norma do gradiente é necessariamente positiva, então o algoritmo converge.

4.1.2 Método de Newton

O método de Newton deriva da expansão de Taylor de segunda ordem:

$$\varepsilon(w(n+1)) = \varepsilon(w(n)) + \nabla^T\varepsilon(w(n))\Delta w(n) + \frac{1}{2}\Delta^T w(n)\nabla^2\varepsilon(w(n))\Delta w(n) \tag{4}$$

Onde $\nabla^T \varepsilon(w(n))$ é o gradiente transposto do custo de aproximação em relação à w e pelo Teorema de Schwarz $\nabla^2 \varepsilon(w(n))$ é a matriz Hessiana m por m simétrica:

$$\nabla^2 \varepsilon(w(n)) = H(w(n)) = \begin{bmatrix} \frac{\partial \varepsilon}{\partial w_1^2} & \frac{\partial \varepsilon}{\partial w_1 \partial w_2} & \cdots & \frac{\partial \varepsilon}{\partial w_1 \partial w_m} \\ \frac{\partial \varepsilon}{\partial w_1 \partial w_2} & \frac{\partial \varepsilon}{\partial w_2^2} & \cdots & \frac{\partial \varepsilon}{\partial w_2 \partial w_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \varepsilon}{\partial w_1 \partial w_m} & \frac{\partial \varepsilon}{\partial w_2 \partial w_m} & \cdots & \frac{\partial \varepsilon}{\partial w_m^2} \end{bmatrix} \tag{5}$$

Então queremos achar a direção de correção de w tal que a variação do custo seja **0**. Derivando em ambos os lados e igualando a zero,

$$\frac{d}{dw(n)} \Delta\varepsilon(w(n)) = \nabla\varepsilon(w(n)) + \frac{1}{2} \nabla^2 \varepsilon(w(n)) \Delta w(n) = 0 \tag{6}$$

Implica que

$$\Delta w(n) = -2\nabla^{-2} \varepsilon(w(n)) \nabla\varepsilon(w(n)) = -2H^{-1}(w(n)) \nabla\varepsilon(w(n)) \tag{7}$$

Apesar de convergir mais rapidamente e com maior precisão que a descida do gradiente, por fazer uso da forma quadrática da função de custo, o método de Newton tem algumas limitações. Primeiro, o custo de aproximação deve ser uma função de classe C^2, ou seja, que admite todas as derivadas parciais de segunda ordem, e segundo, a matriz hessiana deve ser definida positivamente a cada iteração para admitir sua inversa.

4.1.3 Algoritmo de Retropropagação dos Erros

O algoritmo de Retropropagação (RUMELHART, HINTON e WILLIAMS, 1986) ou "BP" consiste de duas etapas: a primeira de propagação do sinal de entrada e a segunda de retro-propagação do sinal de erro na saída. Na primeira etapa, uma observação das variáveis de entrada é apresentada à rede por ciclo de treinamento. Cada observação de treinamento consiste em um vetor $\{x_i\}_{i=1}^m$, $x_i \in \mathbb{R}^n$.

A arquitetura mais simples de uma rede neural é denominada *Feedforward Multilayer Perceptron*, mostrada na Figura 4.2. A arquitetura tem uma camada de sinais de entrada $(x_1, x_2, \dots, x_n)$, uma camada de neurônios intermediária (também conhecida como camada oculta) conectada pelo conjunto de sinapses $\{w_{ij}\}_{i,j=1}^{n,L}$. E uma camada de saída, aqui composta por apenas um neurônio e conectada pelas sinapses $\{s_i\}_{i=1}^{L}$.

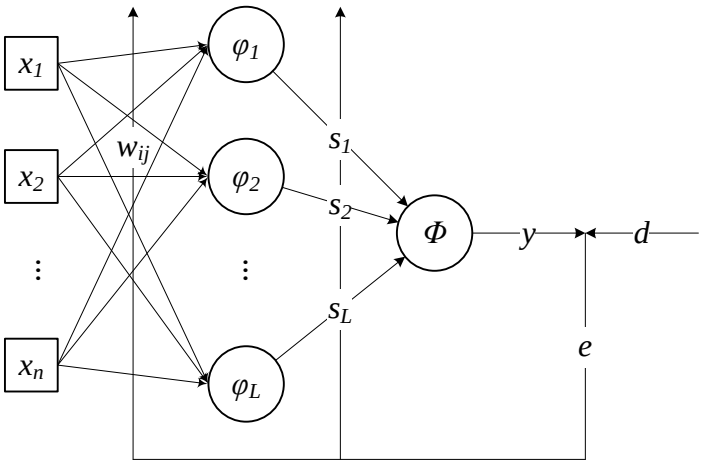

Figura 4.2 Arquitetura de uma Feedforward MLP retropropagada simples

Os sinais $(x_1, x_2, \dots, x_n)$ recebidos são propagados através da rede, sofrendo modulações por meio das sinapses e das funções de transferência no interior dos neurônios $\{\varphi_i\}_{i=1}^{L}$ e ϕ.

A cada ciclo de aprendizagem são feitas correções nas variáveis livres do sistema (sinapses) com o objetivo de realizar a melhor aproximação possível entre o sinal de saída y e o valor observado real d. É importante notar que tanto na figura acima quanto a notação usada a seguir, são consideradas as variáveis sob um estado do sistema no instante t.

Seguindo a evolução da propagação dos sinais, temos primeiramente que x_i é combinado linearmente com a matriz de pesos sinápticos $W \in \mathbb{R}^{L,n}$, produzindo os campos locais induzidos $v \in \mathbb{R}^{L}$: $W * x = v$. A seguir, o campo local induzido é transformado por uma determinada função de transferência $\varphi_i = \varphi_i(v_i)$. As funções corriqueiramente usadas na literatura são a logística e a tangente hiperbólica dadas abaixo:

$$\varphi(x) = \log(x) = \frac{1}{(1 + e^{-x})}$$

$$\varphi(x) = \tanh(x) = \frac{(e^{x} - e^{-x})}{(e^{x} + e^{-x})}$$

Mais uma vez os sinais são combinados linearmente pelas sinapses no campo local induzido $V = S' * \varphi$ do neurônio de saída. E mais uma transformações não-linear ocorre com $y = \phi(V)$.

Após a propagação do sinal do instante t, a energia do erro (ε) equivale à soma dos quadrados dos erros instantâneos de l neurônios de saída. No entanto, em um perceptron com $l = 1$ (como na Figura 4.2), a energia do erro equivale a:

$$\varepsilon = \frac{1}{2} e^{2}$$

O erro instantâneo e é a diferença entre a saída estimada y e o valor esperado pela supervisão da aprendizagem d: $e = y - d$. Objetiva-se que $\varepsilon(t + 1) \leq \varepsilon(t)$.

Para corrigir as sinapses conectadas à camada de saída iterativamente, é necessário descobrir a direção de decrescimento da função ε em relação às sinapses s. Para isso aplicamos o método de Descida do Gradiente. O vetor gradiente $\frac{\partial \varepsilon}{\partial s}$ aponta a direção de maior crescimento de ε em relação à s. Dessa forma,

$$\frac{\partial \varepsilon}{\partial s_l} = e \frac{\partial e}{\partial s_l} = e \frac{\partial (y - d)}{\partial s_l} = e \frac{\partial \big(\phi(V)\big)}{\partial s_l} = e \frac{\partial \left(\phi\big(\sum_{j=1}^{L} s_j \varphi_j\big)\right)}{\partial s_l}$$

Já que d é constante. Para a conexão pertinente na qual $j = l$,

$$\frac{\partial \left(\phi\big(\sum_{j=1}^{L} s_j \varphi_j\big)\right)}{\partial s_l} = \frac{\partial \big(\phi(s_l \varphi_l)\big)}{\partial s_l} = \phi'(s_l \varphi_l) \varphi_l$$

Usando a constante η para regular o tamanho do passo dado em direção à correção, chegamos à correção iterada dos pesos sinápticos de saída com o oposto do gradiente (ver 4.1.1):

$$s_l(t + 1) = s_l(t) - \eta \frac{\partial \varepsilon}{\partial s_l} = s(t) - \eta e \phi'(s_l \varphi_l) \varphi_l$$

O mesmo se aplica aos neurônios da camada oculta, mantendo as sinapses de saída constantes temos o gradiente do erro para a sinapse oculta $w_{i,j}$:

$$\frac{\partial \varepsilon}{\partial w_{i,j}} = e\frac{\partial\big(\phi(V)\big)}{\partial w_{i,j}} = e\phi'(V)\frac{\partial V}{\partial w_{i,j}} = e\phi'(V)\frac{\partial[\sum_{l=1}^{L} s_l\varphi_l]}{\partial w_{i,j}}$$

Tomando apenas o neurônio j que conecta à entrada x_i:

$$\frac{\partial[\sum_{l=1}^{L} s_l\varphi_l]}{\partial w_{i,j}} = \frac{\partial s_j\varphi_j}{\partial w_{i,j}} = s_j\frac{\partial \varphi_j}{\partial w_{i,j}}$$

$$\frac{\partial \varphi_j}{\partial w_{i,j}} = \frac{\partial \varphi_j\big(v_j\big)}{\partial w_{i,j}} = \varphi'_j\big(v_j\big)\frac{\partial v_j}{\partial w_{i,j}}$$

$$\frac{\partial v_j}{\partial w_{i,j}} = \frac{\partial\big[\sum_{i=1}^{n} x_i w_{i,j}\big]}{\partial w_{i,j}} = x_i$$

Juntando os termos,

$$\frac{\partial \varepsilon}{\partial w_{i,j}} = e\phi'(V)s_j\varphi'_j\big(v_j\big)x_i$$

De tal forma que

$$w(t+1) = w(t) - \eta\frac{\partial \varepsilon}{\partial w_{i,j}} = w(t) - \eta e\phi'(V)s_j\varphi'_j\big(v_j\big)x_i$$

Em (HAYKIN, 2001) há a mesma dedução da Retropropagação por Descida do Gradiente para o caso geral, de um *Feedforward Multilayer Perceptron* com mais de uma camada oculta.

4.1.4 Gradiente Conjugado Escalado (SCG)

A limitação do BP está tanto em sua natureza estritamente linear de aproximação da função, quanto no tamanho do passo de correção que é fixo ao longo do tempo (MOLLER, 1993).

O SCG nasceu com a proposta de evitar a entrada de parâmetros altamente dependente do usuário, tal como o tamanho do passo de correção η mostrado em 4.1.3. Além disso, partiu

do ponto de vista da comunidade de pesquisa de análise numérica, por isso, avaliou a necessidade de aplicar algoritmos de otimização para problemas de larga escala já empregados no campo. Em específico, o método de Gradiente Conjugado.

Num estudo comparativo sobre a velocidade de computação e precisão de convergência de 9 algoritmos de aprendizagem com 6 bases de dados, Beale, Hagan e Demuth (2011) observaram que o algoritmo Levemberg-Marquadt tende a ser o mais rápido e convergir a valores mais precisos em problemas com poucos pesos, mas a medida que o tamanho da rede cresce a vantagem se reduz. Além disso, LM tem um desempenho relativamente fraco com problemas de reconhecimento de padrões e os requisitos de memória são maiores que os demais algoritmos. A Retropropagação Resiliente (RP) é o algoritmo mais rápido em problemas de reconhecimento de padrões, mas não tem desempenho satisfatório em aproximação de função.

A Gradiente Conjugado Escalado (SCG) em particular teve um desempenho bom numa grande quantidade de problemas. É quase tão rápido quanto o LM em problemas de aproximação de função, porém mais rápido em redes maiores, e também é quase tão rápido quanto a RP em reconhecimento de padrões. Além disso, requisita uma quantidade modesta de memória.

CAPÍTULO 5 METODOLOGIA

Pode ser dito que as três pesquisas que mais inspiraram a metodologia encontrada aqui foram Peramunetilleke (1997), Mittermayer (2004) e Schumaker e Chen (2009). O primeiro por ser o precursor em FIR e pela comparação feita entre diversos métodos de pesagem comuns em IR, o segundo pela abordagem focada em comunicados ao mercado e o terceiro por experimentar a regressão com os dados textuais.

Neste capítulo será dada uma descrição das notícias publicadas por duas das maiores agências de notícias financeiras do mercado que foram selecionadas como fontes de informação. A seguir há o detalhamento da montagem do banco de dados e dos algoritmos de pré-processamento. Uma explicação da metodologia de análise será feita e por fim os indicadores de desempenho com seus respectivos testes de hipótese.

5.1 Fontes de notícias

Em primeiro lugar, a seleção do mercado norte-americano – e conseqüentemente da língua inglesa – ocorreu pela constatação de que no Brasil o volume de comunicados à imprensa ainda está amadurecendo. Numa coleta recente de comunicados à imprensa da PRNewsWire ao longo de uma semana, o volume de notícias em português era um centésimo do volume em inglês. Como esta pesquisa conta com um volume elevado de dados para que seja possível extrair uma amostra representativa depois de muitas etapas de filtragem, optou-se pelo canal internacional. Além do mais, grande parte das pesquisas na área estudou ou o mercado norte-americano ou o asiático, então isso também facilitaria a comparação dos resultados.

Foram selecionadas 2 fontes de notícias neste trabalho, PRNewsWire e Businesswire, ambas publicadoras de comunicados à imprensa. De acordo com Mittermayer (2004), o interesse em trabalhar com *press releases* está no fato de serem o hub primário na disseminação de um fato relevante ao mercado. Além disso, cada uma das duas agências são responsáveis por metade do mercado de *press releases*.

A Business Wire, fundada em 1961, é uma empresa que dissemina a publicação de artigos de notícias de empresas e organizações às mídias de notícias, mercados financeiros, investidores, web sites de informação, bases de dados e outros públicos. A empresa possui acordos de fornecimento com agências de notícias diversas, incluindo a *Associated Press*, *Agence France-Presse*, *Bloomberg*, *Dow Jones*, *Reuters*, *Thomson One*, e outras 60 agências de notícias regionais, para entregar o conteúdo diretamente em seus sistemas editoriais de redação. Foi o primeiro serviço deste tipo a colocar as notícias on-line, lançando o site da empresa em maio de 1995.

A PR Newswire começou em 1954 como fornecedora de serviços contratada por empresas e agências para enviar *press releases* para a mídia. Com sede em Nova York, a PR Newswire tem escritórios em 16 países e envia notícias para publicação em 135 países. Ela também tem acordos com agências de notícias mundiais, como *Bloomberg*, *Associated Press*, *Dow Jones* e *Reuters Thomson* para enviar informações diretamente para redações.

De acordo com Thomas, 2003, PR Newswire e Business Wire são responsáveis por 30% do volume online de notícias de todas as ações, e 19% das notícias do quinto de ações mais noticiadas. Isso mostra que as agências de *press releases* têm uma distribuição de publicação mais uniforme dentre as ações pesquisadas pelo autor.

Tabela 5.1 Volume de Notícias por Fonte (retirado de THOMAS, 2003)

Fonte	Todas Ações	Quinto Maior
Reuters	21.66%	28.59%
PR Newswire	17.35%	11.02%
Business Wire	13.22%	7.82%
CBS Marketwatch	8.20%	12.00%
Thestreet.com	4.19%	6.48%
CCBN	3.79%	0.86%
EDGAR	3.77%	0.98%
Briefing.com	3.43%	1.79%
Thomsonfn.com	2.86%	0.71%
ON24	2.85%	4.25%
Wall Street Journal	2.07%	3.20%
Forbes.com	2.02%	3.33%
Yahoo!	1.81%	2.71%
Associated Press	1.73%	2.59%
First Call	1.47%	0.34%
Internet Wire	1.40%	1.43%

Um ponto não considerado, mas certamente mais difícil de ser avaliado, é a proporção de notícias veiculadas pelas demais agências que replicam os comunicados distribuídos pela PR Newswire ou pela Business Wire.

5.2 Arquitetura do sistema

A arquitetura do sistema é descrita a seguir, com cada subseção à essa cobrindo um módulo do sistema: estruturação do banco de dado de notícias, pré-processamento e inferência.

5.2.1 Base de Dados de Notícias

A ferramenta desenvolvida segue a mesma estrutura de IR, com 3 módulos seqüenciais de coleta de dados, pré-processamento e indução. O módulo de coleta de notícias foi programado em PHP+Curl e consiste de dois *web crawlers*[5], um específico para cada fonte de artigos (PRnewswire e BusinessWire).

PHP é uma linguagem de programação livre de paradigma procedural / orientada a objetos (POO), utilizada para gerar conteúdo dinâmico na WWW. Curl é uma biblioteca

[5] *Web Crawler* ou *Spider* são algoritmos de coleta automática de dados na internet. Podem ser usados para vários propósitos, desde fornecimento de dados para uma ferramenta de busca (exemplo o GoogleBot) até manutenção de *websites*.

suportada pelo PHP que permite a conexão e comunicação com muitos tipos diferentes de servidores e protocolos (ARNTZEN, BAKKEN, *et al.*, 2011).

Cada *web crawler* segue a mesma rotina de instruções de *data scraping*, iniciadas no portal de notícias de cada fonte. *Data scraping* é uma técnica de extração de dados a partir de *output* legível por humanos, neste caso, o código-fonte das páginas web. A PRnewswire publica a listagem completa de artigos em www.prnewswire.com/news-releases/news-releases-list/, onde mantém uma base online atualizada em tempo real com artigos de mais de 5 meses. A BusinessWire mantém uma base semelhante com até 1 mês de histórico em www.businesswire.com/portal/site/home/news/.

Cada *crawler* faz paginação dos portais enviando uma requisição de leitura da URL e armazena o conteúdo do código (*script*) numa (variável) *string* temporária. As páginas dos portais seguem um comportamento de consulta estático, ao contrário, por exemplo, do que acontece com o Google News (www.news.google.com), que retorna na URL o código – aparentemente aleatório – da requisição de consulta. Até o momento de redação deste trabalho, a PRnewswire organiza as notícias de maneira simples, concatenando ao prefixo de caminho (*path*) a string de consulta (*query string*) '`/?page=`' mais o índice da página atual. A BusinessWire concatena uma string arbitrária do tipo '`/template.PAGE/news/?javax.portlet.prp_ccf123a93466ea4c882a06a9149550fd_ndmHsc=v2`' ao prefixo, mais '`*A`' e mil vezes a data inicial da consulta em formato Unix, mais '`*G`' e o índice da página atual.

Com o conteúdo de cada página armazenada, os *web crawlers* identificam os links dos artigos com base na estrutura de HTML *tags* internas do *script*, que também são estáticos. O uso de Regular Expressions (*RegEx*) se faz necessário, e o suporte completo à gramática de *RegEx* oferecida por linguagens web como PHP ou JavaScript aliado às bibliotecas nativas (*in-built*) apontam para o uso delas na implementação.

Cada *link* que combina com o formato *RegEx* pré-definido para um artigo (i.e. cada *match*) é visitado e o *script* da notícia é extraído. O Quadro 5-1 abaixo mostra um exemplo da notícia "Microsoft Launches Office 365 Globally", publicada na PR Newswire em 28/junho/2011. O exemplo ilustra a presença de HTML tags (e.g. `<p>` ou `<span>`) e da data capturada referente à notícia.

```
      <p><span class="xn-location">NEW YORK</span>, <span class="xn-
chron">June 28, 2011</span> /PRNewswire/ --<b> </b>Today, at media events
around the world, Microsoft Corp. announced the availability of <a
onClick=linkOnClick(this) target='_blank'
href="http://office365.microsoft.com/">Microsoft Office 365</a>, the
company's newest cloud service. Office 365 is now available in 40 markets,
and it brings together Microsoft Office, Microsoft SharePoint Online,
Microsoft Exchange Online and Microsoft Lync Online in an always-up-to-date
cloud service, at a predictable monthly subscription.</p>
```

Quadro 5-1 Extrato de código-fonte de um press release na PR Newswire

O conteúdo extraído é convertido em UTF-8 (*8-bit Unicode Transformation Format*) para unificar a codificação das *strings*. A seguir, a primeira camada de ruído é removida, com remoção das HTML *tags* (estilos, objetos, applets, frames, etc.), espaços em branco, cabeçalhos e rodapés de origem específica do site. As datas e manchetes também são processadas para armazenamento em formato adequado.

Junto com o conteúdo, extraem-se também os meta-dados, isto é, a data e horário da publicação e o título ou manchete (*headline*) da notícia. Todas essas informações são gravadas no banco de dados de notícias, aqui denominado BDnews. O documento recuperado e limpo é salvo em um arquivo como mostrado abaixo, resultado da extração do quadro anterior.

```
      Microsoft Launches Office 365 Globally

      28/Jun/2011 @ 10:00

      http://www.prnewswire.com/news-releases/microsoft-launches-office-
365-globally-124644918.html

      PR

      NEW YORK, June 28, 2011 /PRNewswire/ -- Today, at media events around
the world, Microsoft Corp. announced the availability of Microsoft Office
365, the company's newest cloud service. Office 365 is now available in 40
markets, and it brings together Microsoft Office, Microsoft SharePoint
Online, Microsoft Exchange Online and Microsoft Lync Online in an always-
up-to-date cloud service, at a predictable monthly subscription.
```

Quadro 5-2 Extrato de Documento armazenado depois da filtragem

Para uma aplicação online de um sistema de trading, o ideal é montar o banco de dados relacional (RDBMS), como MySQL ou SQLserver, pela velocidade e flexibilidades de

consulta. Neste estudo cada notícia foi gravada num arquivo de texto (.txt) com os meta-dados no cabeçalho, a escolha feita com base na maior velocidade de implementação em detrimento dos demais critérios. O BDnews deste estudo contém 109.905 notícias continuamente publicadas entre [2011-06-16@16:13] e [2011-09-20@23:56] por ambas agências de notícia.

5.2.2 Pré-Processamento

A Figura 5.1 mostra o fluxograma do sistema de pré-processamento das notícias. O sistema é alimentado pelo banco de dados com notícias mineradas (no presente trabalho denominado BDnews), e pelas cotações dos ativos analisados.

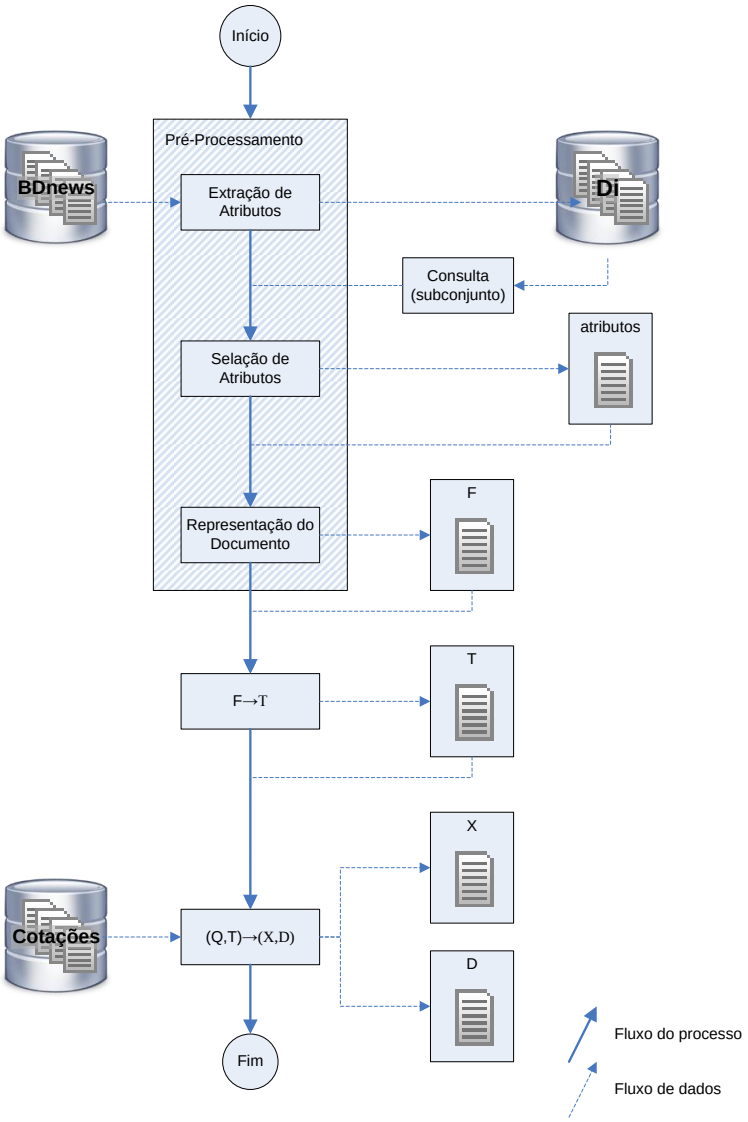

Figura 5.1 Fluxograma do pré-processamento

Como discutido no capítulo 2.6, o pré-processamento é iniciado com a Extração de Atributos, que percorre cada notícia minerada para listar as palavras e suas contagens. O Pseudocódigo 5-1 mostra exatamente o interior de processo:

```
ALGORITMO Extração de Atributos
pp[] ← palavras_de_parada()
pi[] ← palavras_indesejadas()
INICIO
para i ← 1 até |BDnews|
      docstring ← carrega_documento()
      docstring ← remove_caracteres_indesejados(docstring)
      lista[] ← explode(docstring)
      lista[] ← lista[] − (lista[] ∩ pp[])
      lista[] ← radiciacao(lista[])
      para j ← 1 até |lista|
            frequencia(palavra(j))++
      fim para
      salva(freqüência[], /D(i)/freq)
      limpa_memoria(docstring, lista[], frequencia[])
fim para
FIM INICIO
```

Pseudocódigo 5-1 Extração de Atributos

Na notação usada, parênteses são usados para entrada de parâmetros em função, as chaves `[]` indicam vetores ou matrizes e o símbolo ← é o operador de atribuição. Além disso, `|.|` é a função de comprimento, o operador `++` é o incremento unitário e `/dir/` representa diretório acessado.

A função `explode(.)` – comum a qualquer linguagem – quebra as palavras de uma *string* separadas por espaço, tabulação ou nova linha em um vetor de *strings*. A representação `frequencia(palavra(j))` é conhecida como *array* associativo em PHP ou dicionário em VBA e C#, e nada mais é que um vetor indexado por *strings*. A função

`remove_caracteres_indesejados(.)` remove pontuação, símbolos, números e converte caracteres maiúsculos em minúsculos.

Assim, um novo banco de dados paralelo ao BDnews é montado, aqui denominado Di (em referência ao i-ésimo artigo da base), contendo vetores de palavras e suas contagens representando o texto original do documento. Essa redundância é necessária para que seja possível lidar com o volume muito grande de dados. Supondo que a etapa de Extração de Atributos fosse pulada, então uma matriz de completa deveria ser construída, representando todos os documento e palavras para então grande parte da matriz ser podada. É claro que a memória para isso é inviável (cerca de 2^{21} palavras por 2^{17} documentos de 2^{2} `int` equivale a 1 terabyte, neste trabalho). Da mesma forma, adicionar e remover atributos à matriz dinamicamente seria custoso em termos de processamento e implementação.

A consulta à base de vetores é genérica: poderia ser implementada como uma consulta SQL comum numa arquitetura relacional. Para este sistema foi escrito um código simples que retorna interseções de uma listagem de palavras de *input* com o conjunto de vetores disponível. Um exemplo desse processo seria procurar apenas por documentos que contenham os *tickers* de empresas do setor GICS de tecnologia da informação.

O resultado da extração de atributos do documento exemplificado é dado no Quadro 5-3 abaixo, onde se pode ver a contagem das 10 palavras mais freqüentes no documento e o total no cabeçalho.

```
1307
and   45
the   39
offic 33
of    25
to    25
for   23
microsoft   22
a     21
servic      20
avail 9
```

Quadro 5-3 Extrato do vetor de atributos extraídos

A seguir a Seleção de Atributos recebe o subconjunto de vetores de Di consultado e retorna o vetor de palavras do dicionário com maior peso de ocorrência, como mostrado no pseudocódigo abaixo:

```
ALGORITMO Seleção de Atributos
cs ← corte_superior
ci ← corte_inferior
INICIO
para i ← 1 até |BDnews|
      frequencia[] ← carrega(/D(i)/freq)
      para j ← 1 até |frequencia|
            peso_parcial[j] = pesagem(freqüência[j], n)
            peso_total[j] = peso_total[j] + peso_parcial[j]
      fim para
fim para
peso_total[] ← classifica_por_numero(peso_total[])
atributos ← corta_vetor(peso_total, cs, ci)
salva(atributos, /atributos)
FIM INICIO
```

Pseudocódigo 5-2 Seleção de Atributos

Em palavras, tem-se que o algoritmo faz a leitura do dicionário (vetor de contagem de termos) de cada documento e acumula toda informação num só dicionário chamado `peso_total`. A classificação é feita sobre os pesos totais em ordem decrescente para então podar da cauda. A cabeça do dicionário pode também ser cortada no ponto de `corte_superior`, com o objetivo de remover palavras demasiadamente comuns e independentes de contexto. No entanto veremos mais adiante na aplicação (capítulo 6.5 Efeito da Dimensão Léxica) que o corte superior foi desnecessário, e que palavras com posição maior que 30 no dicionário global já começam a ser significativas na classificação dos documentos.

```
the   2157958
and   1808420
of    1320834
to    1165172
in    999433
a     933629
for   876933
new   563434
pr    467488
on    429110
newswir     390174
servic      357615
```

Quadro 5-4 Extrato da seleção de atributos

O Quadro 5-4 acima mostra os 10 primeiros resultados da seleção de atributos dentre todos os documentos, com a respectiva contagem de cada termo.

```
ALGORITMO Representação de Documentos
Fi ← ()
INICIO
atributos[] ← carrega(/atributos)
para i ← 1 até |BDnews|
      frequencia[] ← carrega(/D(i)/freq)
      para j ← 1 até |atributos|
            p = atributos[j]
            se frequencia[p] ≠ Ø, Fi[j] ← frequencia[p]
            senão Fi[j] ← 0;
      fim para
      escreve(Fi[] & nova_linha, /F)
fim para
FIM INICIO
```

Pseudocódigo 5-3 Representação de Documentos

A Representação de Documentos se encarrega de pegar os atributos pré-selecionados pelo algoritmo anterior, e procurar pelas correspondências de palavras na base de documentos com o objetivo de montar a matriz F de freqüência das palavras. A matriz F recebe uma impressão de escrita por documento também por motivos de economia de memória. O operador & denota concatenação.

16/Jun/2011 16:13	40	48	47	19	18	21	17	17	14
16/Jun/2011 16:28	38	38	39	28	16	19	17	13	18
16/Jun/2011 16:53	41	19	13	5	6	25	5	9	7
16/Jun/2011 17:28	37	53	21	17	16	17	15	19	32
16/Jun/2011 17:33	43	25	47	22	10	18	10	19	8
16/Jun/2011 17:33	45	43	34	32	14	16	18	14	8
16/Jun/2011 18:14	47	50	46	25	15	24	11	15	13
17/Jun/2011 03:00	50	41	46	23	23	21	15	17	12
17/Jun/2011 05:30	45	41	21	18	15	14	17	12	11
17/Jun/2011 06:00	40	24	28	17	11	16	8	12	7

Quadro 5-5 Extrato da matriz F de frequencias

Uma parte do resultado final pode ser visualizada acima no Quadro 5-5 acima. Nesta implementação, a matriz recebe nas primeira e segunda coluna os rótulo de data e hora da publicação, para mais tarde F ser alinhada à matriz de cotações Q (ver Figura 5.1).

5.3 Algoritmos de Inferência

A ferramenta de inferência usada será o Perceptron Multicamadas com aprendizagem por Gradiente Conjugado Escalado (SCG).

As redes neurais foram selecionadas por 2 razões: primeiro, são capazes de realizar filtragem e separação de padrões não-lineares sobre os dados. Um exemplo disso é dado em (HAYKIN, 2001), onde o autor discute a separação do problema XOR (Ou-Exclusivo). Sucintamente, este problema consiste em separar automaticamente um padrão onde 0+0=1,

0+1=1, 1+0=0, porém 1+1=0 e não 2, como seria de se esperar de um problema linear. O que isso pode significar no processamento de textos é que relações não-lineares entre palavras podem ser capturadas por meio de uma rede neural.

Segundo, por meio de uma ligeira conversão na arquitetura da rede é possível cambiar entre as tarefas de regressão e classificação. Isso facilita o trabalho de programação, mas principalmente, permite a comparação imediata entre o desempenho de ambas as tarefas frente a um mesmo método, que é uma das propostas desta pesquisa.

O SCG foi escolhido como algoritmo de aprendizagem por convergir mais rápido que diversos outros algoritmos, tais como o *Backpropagation* tradicional (BP), *Conjugate Gradient with Line Search* (CGL) e Broyden-Fletcher-Goldfarb-Shanno quasi-newton (BFGS) (MOLLER, 1991).

Nas aplicações deste trabalho o SCG será usado tanto para regressão quanto classificação, e pelos testes conduzidos se mostrou um algoritmo eficiente (com tempos de convergência muito menores que LM e sem estouro de memória com um elevado número de neurônios ocultos) e confiável (convergências sucessivas a valores de minimização da função-objetivo próximos e número estável de iterações até atingimento do critério de parada). No CAPÍTULO 6 serão dados maiores detalhes sobre os parâmetros avaliados.

5.4 Definição do Modelo

Para um dado mapeamento de entrada e saída $x(t) = \{x_1(t), x_2(t), \dots, x_n(t)\} \rightarrow y(t + \Delta t)$ com Δt sendo o tempo de defasagem da previsão, o modelo de regressão estima a saída pela propagação da MLP, como visto na seção 4.1.3:

$$y(t) = \sum_{l=1}^{L} w_{lk}^{(2)} \varphi_l \left(\sum_{j=0}^{n} w_{jl}^{(1)} x_j(t - \Delta t) \right)$$

Contudo, para classificação preditiva, aplica-se de uma transformação sobre a saída $y_k(t)$ para a classificação sobre a classe k:

$$y_k(t) = \varphi_k \left(\sum_{l=1}^{L} w_{lk}^{(2)} \varphi_l \left(\sum_{j=0}^{n} w_{jl}^{(1)} x_j(t - \Delta t) \right) \right)$$

.

$$C\big(y_k(t)\big) = \max_K(\{y_k(t)\})$$

A função $C(.) = \{1, \dots, K\}$ se traduz na aprendizagem competitiva da rede neural, na qual os neurônios de saída de índices $k = 1 \dots K$ competem entre si na proposta da resposta de maior confiança estatística sob a minimização do erro quadrático. O vencedor ("*winner-takes-all"*) é responsável não ó pela resposta da rede, como também concentra toda a aprendizagem na etapa de retropropagação.

Neste modelo, $x_j(t - \Delta t)$ corresponde ao vetor de contagem de palavras por documento (i.e. $W_{i\diamond}$ como visto em 2.6.2) do instante passado e $y(t)$ é o retorno da ação no instante atual. Contudo, $y_k(t)$ é a resposta da rede sobre a relação do estado observado $x_j(t - \Delta t)$ com a classe k.

5.5 Medidas de Desempenho

A disciplina de IR se desenvolveu de forma altamente empírica (MANNING, RAGHAVAN e SCHÜTZE, 2009). Questões sobre o uso ou não de radiciação de palavras(SOARES, PRATI e MONARD, 2009), seleção de atributos(YANG e PEDERSEN, 1997) ou ferramentas de indução(DUMAIS, PLATT, *et al.*, 1998) foram detalhadamente avaliadas por meio de testes comparativos de desempenho.

A tarefa executada por sistema de IR é avaliada de acordo com a relevância dos documentos retornados pela consulta. É importante observar que relevância não equivale à correspondência dos termos pesquisados com os termos retornados. Exemplos disso são palavras homônimas, como 'laranja' (que tem sentido de cor ou fruta), 'planta' (árvore ou projeto) e 'python' (serpente ou linguagem de programação). Destas palavras, apenas um significado é relevante. Isso reflete no diferencial entre sistemas de busca simples e sistemas baseados em aprendizagem de máquina. A relevância pode ser auferida em escala de avaliação pelo cliente do sistema, que no seu modo mais simples é binária (relevante / não-relevante).

Duas das medidas de relevância mais comuns em IR são a Precisão (*Precision*) e a Revocação (*Recall*). Uma terceira medida, muito comum não só a IR como também à

previsão de séries temporais financeiras é a Acurácia (*Accuracy*). Além dessas 3 medidas de desempenho, será avaliado o Erro Quadrático Médio (*MSE*) das estimativas. E nos estudos envolvendo regressão, o coeficiente de correlação de Pearson (R^2) será avaliado. A acurácia e o R^2 serão estatisticamente testados sob a hipótese nula de que as estimativas geradas pela inferência são aleatórias, e o P-valor desses testes será sistematicamente avaliado.

5.5.1 Matriz de Confusão

Antes de entrar na descrição dos indicadores, e considerando que o trabalho usará extensivamente métodos de classificação, é importante detalhar a Matriz de Confusão. Eventualmente também denominada Matriz de Contingência, apresenta de forma resumida todo o resultado de classificação de um problema.

		y Real (alvo)			
		y_1	y_2	...	y_L
ŷ Estimado (saída)	$\hat{y}_1$	C_{11}	C_{12}	...	C_{1L}
	$\hat{y}_2$	C_{21}	C_{22}	...	C_{2L}
	⋮	⋮	⋮	...	⋮
	$\hat{y}_l$	C_{L1}	C_{L2}	...	C_{LL}

Figura 5.2 Matriz de Contingência genérica

Sejam dois vetores $y, \hat{y} \in \mathbb{B}_{m,L}$, respectivamente de valores reais (ou alvos) e estimativas (ou saídas) de um processo de classificação, onde $\mathbb{B}$ é o conjunto $\{0,1\}$, m indica o número de observações e L o número de classes do problema, tal que $\|y\| = \|\hat{y}\| = 1$. A matriz de confusão $C \in \mathbb{R}_{L,L}$ pode ser composta pelo cruzamento de y com $\hat{y}$, ou simplesmente $C = \hat{y}' * y$.

A diagonal branca da matriz C_{ii}, $1 \leq i \leq L$, indica o número de classificações corretas, e a área hachurada C_{ij}, $i \neq j$, indica os erros de classificação.

Em IR a matriz de contingência é geralmente usada com $k = 2$ (MANNING, RAGHAVAN e SCHÜTZE, 2009) como na representação abaixo (Tabela 5.2), onde é de

interesse do pesquisador ter a resposta da classificação em relação a apenas 1 dada classe alvo, independente do número total de classes do conjunto de documentos.

Tabela 5.2 Tabela de contingência de IR

	Relevante	Não Relevante
Recuperado	Positivo Correto (*pc*)	Positivo Incorreto (*pi*)
Não Recuperado	Negativo Incorreto (*ni*)	Negativo Correto (*nc*)

Todavia, em FIR o número de classes importa já que é a partição de um mesmo contradomínio (e.g. os log-retornos de um ativo). Neste trabalho trabalharemos majoritariamente com $L = 3$, mas também serão feitos alguns estudos com L maior. A seguir veremos sua relação com os principais indicadores de IR.

Daqui em diante usaremos $M_{ij} = \frac{C_{ij}}{m}$ a proporção de classificação realizada. Em relação ao teste de hipóteses, considere $y \in \mathbb{B}_{m,L}$ com L classes equiprováveis, e $\hat{y} \in \mathbb{B}_{m,L}$ uma estimação aleatória dada por um classificador qualquer. Então $P(M_{ij}) = P\left(\hat{Y}_i = \hat{y}_i, Y_j = y_j\right) = P\left(\hat{Y}_i = \hat{y}_i\right)P\left(Y_j = y_j\right) = \hat{p}_i p_j$ por independência, $P(M_{ij}) = 1/L^2$.

Se M_{ij} segue uma distribuição $Bernoulli(\theta = \hat{p}_i p_j)$ com sucesso em $\hat{Y}_i = \hat{y}_i, Y_j = y_j$ e fracasso caso contrário, então sabemos que os momentos $\mathbb{E}(M_{ij}) = \theta$ e $\mathbb{E}^2(M_{ij}) = \theta$.

5.5.2 Acurácia

Alguns dicionários descrevem acurácia como neologismo da língua inglesa, e existe alguma confusão no emprego desse termo. Em sua dissertação, (AROUCK, 2011) defende a troca pelo termo precisão. Contudo neste trabalho, definiremos acurácia como a proporção de vezes que o sinal de resposta é corretamente estimado. Na extrapolação de uma regressão, a acurácia será dada por:

$$acuracia(\hat{y}) = \sum_{i=1}^{m} I\{sgn(\hat{y}_i) = sgn(y_i)\}$$

A função $sgn(.)$ é a função sinal com $sgn(0) = 0$, e I é a função indicadora sobre o conjunto dado. Seguindo a mesma definição, mas considerando a extrapolação de uma classificação, temos:

$$acuracia(\hat{y}) = tr(M) = \frac{\sum_{i=1}^{L} C_{ii}}{m}$$

A acurácia é a proporção do traço da matriz de confusão. Em outras palavras, é a fração da informação recuperada e relevante dentre toda informação disponível:

$$\begin{aligned} acurácia &= \frac{|\{documentos\ relevantes\ recuperados\}|}{|\{total\ de\ documentos\}|} \\ &= P\{recuperados\ \cap\ relevantes\} \end{aligned} \tag{8}$$

De acordo com a Tabela 5.2, podemos reescrever o indicador da seguinte forma:

$$acurácia = \frac{|\{documentos\ corretos\}|}{|\{todos\ os\ documentos\}|} = \frac{(pc + nc)}{(pc + pi + ni + nc)} \tag{9}$$

No entanto, a acurácia não costuma ser usada em IR já que costuma ser enviesada. Em grandes bases de dados existe uma elevada assimetria entre *pc* e *nc*: os negativos corretos são em número muito maior. Isso leva o indicador a tender a 100% mesmo sem retornar nenhum positivo correto para a consulta.

No problema de classificação de notícias financeiras, contudo, a assimetria se mantém baixa conforme a partição feita sobre y. Aqui as classes serão arbitrariamente equiprováveis, e a variância da resposta será controlada por meio de teste de hipótese.

Neste caso, sejam $X_1, X_2 \ldots X_m$ uma amostra aleatória de observações independentes de $X \sim B(1/L)$ com $P(sucesso) = 1/L$ se $X \in tr(C)$ de um classificador aleatório. Então a acurácia observada $\overline{acur} = \frac{1}{m}\sum_{i=1}^{m} X_i$ tem por i.i.d[6],

$$\mathbb{E}(\overline{acur}) = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right] = \frac{1}{m} m\mathbb{E}[X_i] = 1/L$$

Da mesma maneira,

$$\mathbb{V}(\overline{acur}) = \mathbb{V}\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right] = \frac{1}{m^2} m\mathbb{V}[X_i] = \frac{1}{m}\frac{1}{L}\left(1 - \frac{1}{L}\right) = \frac{L-1}{mL^2}$$

[6] Variáveis independentes, identicamente distribuídas (i.i.d.)

Esperamos que a acurácia fosse superior à da classificação aleatória. Para testar isso definimos

$$H_0: acur = 1/L$$

$$H_1: acur > 1/L$$

Como $\sum_{i=1}^{m} X_i \sim Bin(m, 1/L)$, para m elevado pode-se usar a aproximação pela Normal: $\overline{acur} = \frac{1}{m}\sum_{i=1}^{m} X_i \sim N\left(\frac{1}{L}, \frac{L-1}{mL^2}\right)$ sob H_0 (classificador aleatório). Portanto o nível descritivo é

$$\rho = P_{H_0}\left(Z > \frac{\overline{acur} - \frac{1}{L}}{\sqrt{\frac{L-1}{mL^2}}}\right)$$

.Com $Z \sim N(0,1)$.

5.5.3 Precisão e Revocação

As duas medidas de relevância mais comuns em IR são precisão (*precision*) e revocação (*recall*). Precisão é a fração dos documentos recuperados que são relevantes, e revocação é a fração dos documentos relevantes que são recuperados:

$$Precisão = \frac{|\{documentos\ relevantes\ recuperados\}|}{|\{documentos\ recuperados\}|} = P\{relevantes\ |recuperados\} \quad (10)$$

$$Revocação = \frac{|\{documentos\ relevantes\ recuperados\}|}{|\{documentos\ relevantes\}|} = P\{recuperados|relevantes\} \quad (11)$$

Desta forma, de acordo com a tabela de contingência de IR,

$$Precisão = \frac{pc}{(pc + pi)} \quad (12)$$

$$Revocação = \frac{pc}{(pc + ni)} \quad (13)$$

Quando faz uma busca corriqueira, o usuário de internet deseja encontrar rapidamente o resultado pesquisado, com a preocupação de que os primeiros resultados sejam necessariamente corretos. Dessa forma, uma minimização de *pi*, e logo, uma maior taxa de precisão atenderá melhor sua demanda.

Por outro lado, pesquisas médicas, legais ou acadêmicas têm preferência pela minimização de negativos incorretos, já que não se dão ao luxo de perder algum resultado relevante. Assim, uma maior revocação é mais apropriada.

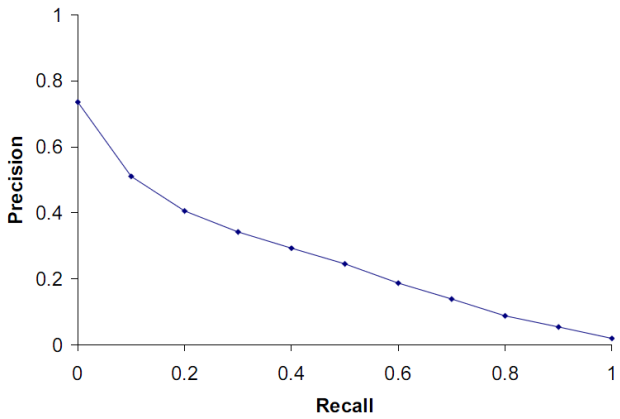

Figura 5.3 Gráfico *P* x *R* com 11 pontos de 50 consultas a um sistema da coleção de dados TREC(MANNING, RAGHAVAN e SCHÜTZE, 2009, p. 160).

A figura acima retirada de (MANNING, RAGHAVAN e SCHÜTZE, 2009, p. 160) ilustra uma relação inversa entre precisão (Precision) e revocação (Recall) comum às pesquisas de IR: à medida que o sistema recupera mais documentos na tentativa de minimizar *ni*, traz ainda mais documentos de *pi*, de forma que precisão e revocação são, de certa forma, inversamente proporcionais.

Contudo, trabalhar com a classificação de mais de 2 classes simultaneamente requer a avaliação de (15) e (16) para cada classe, na forma:

$$Precisão_k = \frac{M_{kk}}{\sum_{i=1}^{L} M_{ki}}$$

$$Revocação_k = \frac{M_{kk}}{\sum_{i=1}^{L} M_{ik}}$$

Dessa forma este trabalho define a precisão e revocação geral do processo de classificação são dadas pela soma ponderada para cada classe, ou,

$$Prec = \frac{1}{L}\sum_{k=1}^{L} Precisão_k$$

$$Revoc = \frac{1}{L}\sum_{k=1}^{L} Revocação_k$$

5.5.4 MSE: Erro Quadrado Médio

O MSE é a forma mais comum de se aferir a diferença entre um estimador e o valor real estimado. É definido como $MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - E[\theta]\right)^2\right]$, e decomposto em na variância do estimador mais seu viés $MSE(\hat{\theta}) = VAR[\hat{\theta}] + [Bias(\hat{\theta},\theta)]^2$ (BOLFARINE e SANDOVAL, 2010). Dois modelos podem ser comparados de acordo com o desempenho em minimizar o MSE sobre um mesmo conjunto de dados. O MSE tem a característica de sobrecarregar os erros de *outliers*, o que em algumas aplicações é indesejável. Para tanto, utiliza-se outras funções tal como o MAE.

5.5.5 Correlação

O coeficiente de correlação linear de Pearson mede o grau de linearidade da variação cruzada de duas variáveis. Sejam X e Y duas variáveis aleatórias, cuja correlação equivale a

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}}$$

Na comparação das estimativas de uma regressão com os valores reais é comum ter a notação trocada de x para $\hat{y}$, a estimativa, e manter y como a variável explicada. A partir daí usa-se o seguinte estimador para o coeficiente de correlação amostral:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Em previsão de séries temporais, espera-se atingir uma correlação amostral próxima de um e estatisticamente distante de zero. Para o teste de hipóteses $H_0: \rho = 0$ contra $H_1: \rho \neq 0$ usa-se a estatística $t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$ sob a hipótese nula..

5.5.6 Classificador Aleatório, Solução Trivial e Classes Equiprováveis

Para um problema com $\mathcal{C}_k$, $1 \leq k \leq L \in \mathbb{N}^*$, classes quaisquer. Definimos um classificador aleatório $\mathcal{F}(X)$ de $X_{mxn} \rightarrow \hat{y}_{mxK}$, $\hat{y}$ binário, é aquele que $P(\mathcal{F}(X_i) = y_i) = {}^{1}/_{L}$, qualquer que seja a partição sobre y.

O teste de hipóteses é conduzido neste trabalho sobre a hipótese nula H_0 de que o classificador foi aleatório, ou informalmente, a classificação das notícias foi um "chute de sorte". Nesse caso, cada classe tem a mesma probabilidade de ser selecionada, como exposto acima.

No entanto quando se usa classes não equiprováveis um novo problema surge para ser testado, nomeado solução trivial. Muitos classificadores são propensos a "apostar" em uma só classe quando não possuem informação suficiente para discriminar as observações dentre as classes desejadas. Mais que isso, um classificador pode se utilizar de viés para "apostar" todas as estimações na classe mais observada, dessa forma reduzindo seu custo de erro de treinamento. Máquinas de aprendizagem baseadas em gradiente podem ser mais propensas a esse tipo de comportamento, mas não posso garantir que uma SVM não se comporte de forma semelhante.

A importância de empregar classes equiprováveis está na eliminação do segundo teste de hipóteses, restando a hipótese de aleatoriedade.

CAPÍTULO 6 ANÁLISE DOS EXPERIMENTOS

Primeiramente será mostrada uma compilação descritiva da base de notícias, seguida pelo pela documentação detalhada do primeiro modelo de previsão com base puramente em notícias. Depois serão realizados diversos experimentos de sensibilidade dos modelos aos parâmetros que cercam a metodologia e finalmente a previsão será estendida ao longo do tempo.

6.1 Base de Dados

A base de notícias contém 109.904 artigos continuamente coletados entre 2011-06-16@16:13 e 2011-09-20@23:56, equivalente a 96 dias corridos. É composta por um dicionário com 556.896 raízes de palavras distintas. A Figura 6.1 mostra a ocorrência de uma pequena amostra das raízes que compõe o dicionário, plotando-se o log da posição dos termos ordenados por ocorrência contra o log de suas ocorrências.

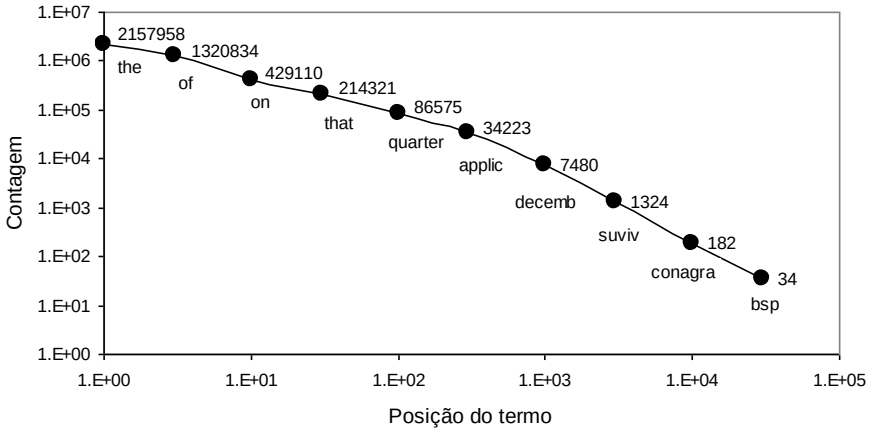

Figura 6.1 Contagem de termos radicados no dicionário universal

Como é de se esperar, artigos e preposições (“the”, “of”, “on”, “that”) tomam o topo das posições no dicionário, seguidos por substantivos e adjetivo comuns (“quarter”, “application”, “december”). Nomes próprios de pequenas empresas, produtos de nicho, sites periféricos e palavras raras se concentram na cauda da distribuição de freqüências (“conagra”, “bsp”). O Apêndice A contém uma lista das raízes mais freqüentes.

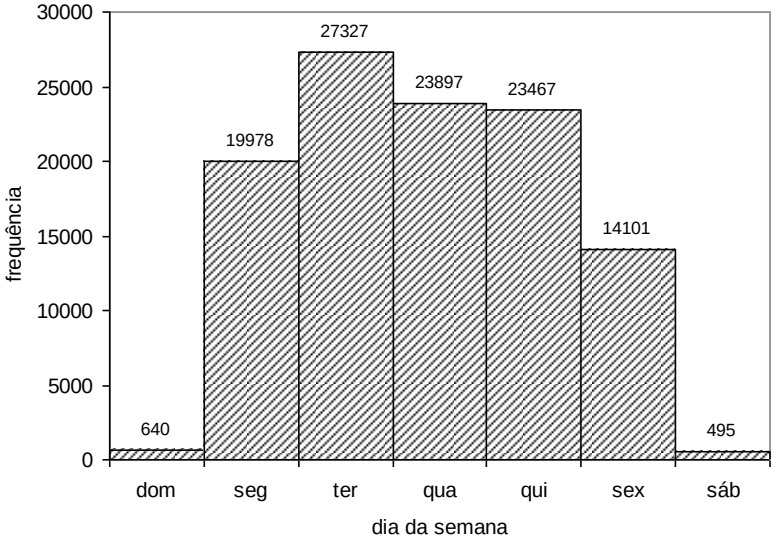

Figura 6.2 Publicação total de comunicados ao longo da semana

Também como se espera, grande parte dos anúncios ocorre nos dias úteis, mas há uma concentração significativa entre terça e quinta feira (Figura 6.2). Os horários de pico de informação ocorrem por volta da abertura do mercado, no entanto é interessante notar como a maior parte é publicada após a abertura, e não antes, como rege a política de divulgação das empresas. Aparentemente isso se deve à maior parte das informações publicadas não se enquadrar na política, ou por serem publicações de baixa relevância, ou ainda, relativas à pequenas empresas não expostas ao riscos de RI.

Um gráfico semelhante ao da Figura 6.3 é encontrado no trabalho de Thomas (2003). Na distribuição do número de publicações da base a cada hora do dia, vemos a maior concentração ocorrendo entre as 9:00 e 10:00 da manhã, acerca da abertura do mercado (às 9:30). Há também uma densidade maior ocorrendo às 17:00, uma hora após o fechamento.

Contudo, o intervalo de negociação contínua ainda contém muitas notícias, o que beneficia os estudos de impacto delas no intraday (em alta freqüência), como é este o caso.

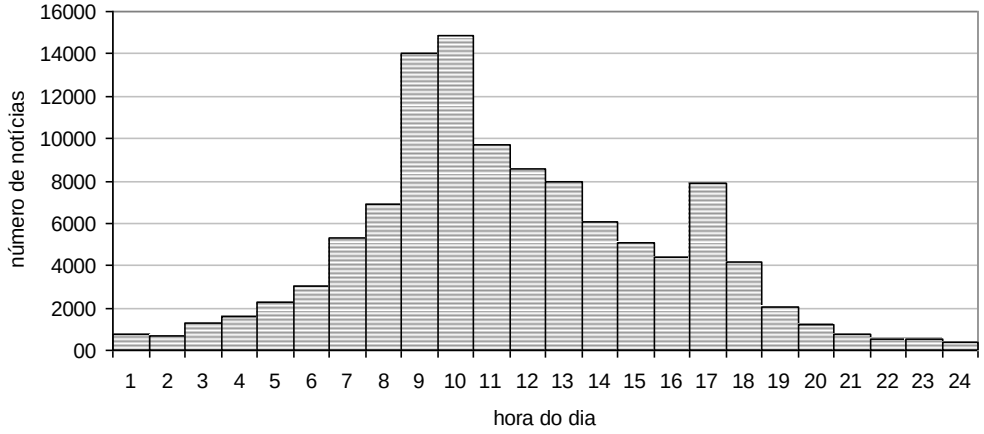

Figura 6.3 Publicação total de comunicados ao longo do dia

Abaixo temos a distribuição pontual dos 22 horários de maior publicação no período estudado. Os instantes de maior publicação dos comunicados ao mercado ocorrem às 8:00 e às 9:00 da manhã em ponto, logo antes das bolsas norte-americanas abrirem o pregão. Certamente há uma predileção pelo agendamento das publicações para esses 2 horários.

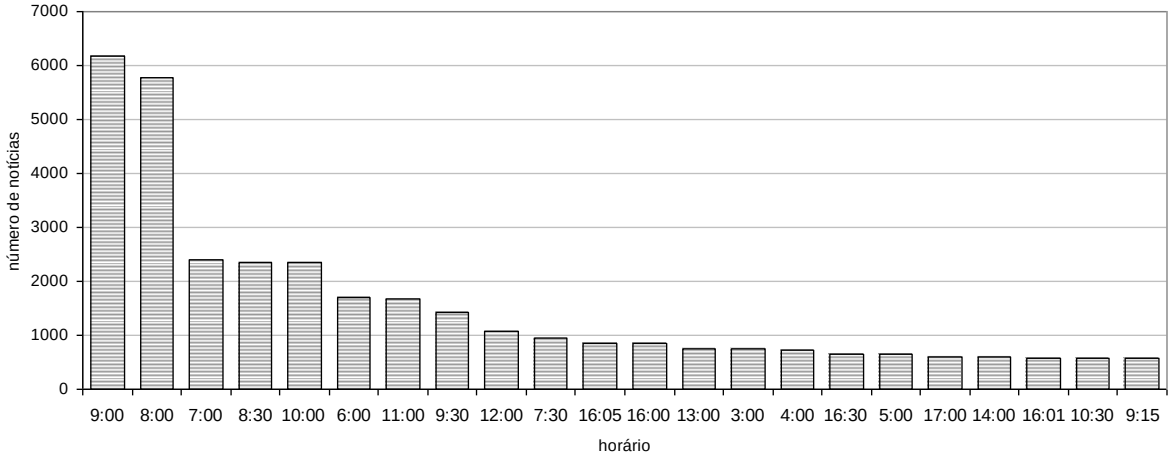

Figura 6.4 Instantes de maior publicação

Considerando a política de divulgação das empresas de capital aberto (ver a seção 3.2.2 Política de Divulgação) e o horário do pregão (das 9:30 às 14:00), é justificado o fato de

que os momentos de maior concentração de publicação ocorrem em horários agendados fora do horário de pregão. Dentro do horário do pregão, a maioria das publicações ocorre gradativamente às 10:00, 11:00, 12:00, 13:00 e 14:00.

6.2 Primeira Classificação

O primeiro teste com o modelo desenvolvido foi sobre a tarefa de classificar o movimento do índice setorial GICS de tecnologia da informação (Bloomberg: S5INFT) entre as mesmas categorias "Up", "Steady" e "Down" utilizadas por Wüthrich e Peramunetilleke (1998). O índice setorial foi selecionado por ter sido apontado por Schumaker e Chen (2009) como o melhor objeto a estimar em termos de acurácia direcional, que é a principal medida de desempenho aqui avaliada. Segundo os autores, a previsão de ações isoladas e a previsão do índice de mercado não apresentaram acurácia tão elevada, provavelmente em função da falta e do excesso de informação, respectivamente.

Da base inicial com 109.904 artigos, foram filtrados 12.102 pela extração de atributos que continham pelo menos uma referência ao nome ou ao *ticker* de alguma das empresas listadas no setor de TI. Como diversos trabalhos anteriores trabalharam com o retorno (GIDÓFALVI, 2001) ou diferença de preço (SCHUMAKER e CHEN, 2009) entre o momento de publicação da notícia e +20 minutos após, manteve-se este mesmo o padrão de alinhamento. Além disso, foi adotado o procedimento de Schumaker e Chen de coletar apenas artigos entre 10:30am e 3:40pm. Apesar de o pregão da NYSE e NASDAQ iniciarem às 9:30am, notícias publicadas fora do horário de pregão podem ter uma janela de impacto incerto na abertura, e o período de 1 hora no início permite que os preços se ajustem e que a volatilidade reduza. Dessa forma, o número de documentos filtrados caiu para 2.520, número semelhante ao usado por outros autores (vide Tabela 6.1).

Tabela 6.1 Coleções de artigos em estudos

# Artigos	
Artigos no estudo de Mittermeyer	6,602
Artigos no estudo de Gidofalvi	5,500
Artigos no estudo de Schumaker	9,211
Artigos filtrados no estudo de Schumaker	2,809
Artigos neste estudo	12,102
Artigos filtrados neste estudo	2,520

Na seleção de atributos foi usado o método de poda de dicionário entre 0% a 0,56% dos termos de maior contagem, resultando num subconjunto de 1.000 radiciações de Porter distintas. A matriz de freqüências F foi então representada com as contagens de termos por documento mais o viés (1.00). As notícias publicadas no mesmo minuto foram agrupadas por representarem um volume de informação concomitante para um mesmo retorno da ação no tempo. Os demais autores não aplicaram esse agrupamento.

Os retornos do índice de TI alinhados com o momento de publicação das notícias +20 min. foram particionados em 3 classes equiprováveis, como mostrado na Tabela 6.2:

Tabela 6.2 Partição do intervalo em Classes

Classe	Intervalo	P({intervalo})
"DOWN"	(-INF ; -7.01e-4)	1/3
"STEADY"	[-7.01e-4 ; 9.05e-4)	1/3
"UP"	[9.05e-4 ; +INF)	1/3

A seguir a inferência foi realizada com um MLP alimentado para frente com algoritmo de aprendizagem SCG, arquitetura 1001-300-3 (número de sinais de entrada – número de neurônios ocultos – numero de neurônios de saída) e tangente hiperbólica em ambas as camadas de transferência. O critério de parada adotado foi de validação cruzada com máximo de 30 falhas consecutivas no decremento do MSE de validação, sobre uma amostragem de treino (70% = 1764), validação (15% = 378) e teste (15% = 378) aleatória.

O elevado número de neurônios ocultos inicialmente escolhido remete ao trabalho de Wüthrich e Permunetilleke (1998) que usaram 211 para um *bag of words* com 423 termos de entrada.

A Tabela 6.3 abaixo resume os parâmetros selecionados dentre todos que a metodologia abordada compreende:

Tabela 6.3 Parametrização do problema inicial

Parâmetros	
Filtro de documentos	TI stocks
Seleção de atributos por	T
Δt agrupamento	1min
Overnight	não
Pesagem	T
Poda de Atributos	1000
Normalização	n/a
Viés	1
y saída	S5INFT
y transformação	retorno líquido +20min
Classes	3 classes 1/3 cada
Método de inferência	mlp
Função de saída	tangente hiperbólica
Algoritmo	SCG
Neurônios ocultos	300
Amostra de treinamento	70% (1764)
Amostra de validação	15% (378)
Amostra de teste	15% (378)
Máximo falhas	30
Número de simulações	30

Na notação adotada para a tabela, T indica a contagem de termos radicados. Overnight indica se notícias publicadas fora do horário do pregão são amostradas.

Foram realizadas 30 simulações, uma vez que a solução de convergência das redes neurais está fortemente relacionada com a inicialização aleatória dos pesos sinápticos. Nelas, a melhor taxa de aprendizagem ocorreu em média a 40,2 iterações do inicio da aprendizagem (escapando 30 ciclos depois a 70,2), num tempo médio de 46,88 segundos. As matrizes de confusão médias são mostradas na Figura 6.5.

Treinamento		Real (alvo) 1	2	3	marginal
Estimado (saída)	1	278.9 (15.8%)	144.5 (8.2%)	164.1 (9.3%)	587.6 (33.3%) 47.5% 52.5%
	2	142.1 (8.1%)	306.5 (17.4%)	140.4 (8.0%)	589.0 (33.4%) 24.1% 75.9%
	3	181.8 (10.3%)	155.6 (8.8%)	250.1 (14.2%)	578.4 (33.3%) 30.9% 69.1%
	marginal	602.8 (34.2%) 46.3% 53.7%	606.6 (34.4%) 23.8% 76.2%	554.7 (31.4%) 29.6% 70.4%	1764 (100%) 47.4% 52.6%

(a)

Validação		Real (alvo) 1	2	3	marginal
Estimado (saída)	1	49.6 (13.1%)	33.3 (8.8%)	41.5 (11.0%)	124.5 (32.9%) 39.9% 60.1%
	2	33.8 (9.0%)	62.1 (16.4%)	30.8 (8.1%)	126.7 (33.5%) 26.7% 73.3%
	3	43.1 (11.4%)	37.5 (9.9%)	46.2 (12.2%)	126.8 (33.5%) 34.0% 66.0%
	marginal	126.5 (33.5%) 39.2% 60.8%	132.9 (35.2%) 25.1% 74.9%	118.6 (31.4%) 35.0% 65.0%	378 (100%) 41.8% 58.2%

(b)

Teste		Real (alvo) 1	2	3	marginal
Estimado (saída)	1	51.6 (13.7%)	33.8 (8.9%)	42.6 (11.3%)	128.0 (33.9%) 40.3% 59.7%
	2	31.9 (8.4%)	60.4 (16.0%)	31.9 (8.4%)	124.3 (32.9%) 25.7% 74.3%
	3	47.3 (12.5%)	35.4 (9.4%)	43.1 (11.4%)	125.8 (33.3%) 37.6% 62.4%
	marginal	130.8 (34.6%) 39.4% 60.6%	129.6 (34.3%) 26.1% 73.9%	117.6 (31.1%) 36.2% 63.8%	378 (100%) 41.0% 59.0%

(c)

Total		Real (alvo) 1	2	3	marginal
Estimado (saída)	1	380.1 (15.1%)	211.7 (8.4%)	248.2 (9.9%)	840.0 (33.3%) 45.3% 54.8%
	2	207.2 (8.2%)	429.0 (17.0%)	203.2 (8.1%)	840.0 (33.3%) 24.7% 75.3%
	3	272.2 (10.8%)	228.4 (9.1%)	339.4 (13.5%)	840.0 (33.3%) 32.4% 67.6%
	marginal	860.1 (34.1%) 44.2% 55.8%	869.0 (34.5%) 24.4% 75.6%	790.8 (31.4%) 31.4% 68.6%	2520 (100%) 45.6% 54.4%

(d)

Figura 6.5 Matrizes de Confusão de (a) Treinamento (b) Validação (c) Teste e (d) Total

Os índices 1, 2 e 3 referenciam as classes “DOWN”, “STEADY” e “UP” respectivamente. As células centrais contêm a contagem média de classificações nas 30 simulações e a porcentagem em relação ao total, e as células exteriores contém mais a porcentagem de acertos e erros das marginais. Observe que a meta é obter mais de 33% de acerto. Em comparação com a matriz de confusão de hipótese nula (ver o capítulo 5.5 Medidas de Desempenho para maiores informações sobre matrizes de confusão e testes de hipótese empregados), as matrizes acima concentram maior densidade na diagonal principal, indicando sucesso na tarefa de classificação. A Tabela 6.4 a seguir resume os principais medidas de desempenho de classificação usadas em IR.

Tabela 6.4 Desempenho de Classificação

Resultados	
melhor iteração	40.2
tempo (s)	46.88
mse treino	0.2091
mse validação	0.2198
mse teste	0.2204
s teste	0.0073
acurácia treino	0.4736
acurácia validação	0.4178
acurácia teste	0.4103
s teste	0.0066
z	3.17085
p-valor	0.08%
-z	2.89900
-z p-valor	0.19%
+z	3.44269
+z p-valor	0.03%
precisão treino	0.4730
precisão validação	0.4164
precisão teste	0.4090
s teste	0.0087
revocação treino	0.4736
revocação validação	0.4177
revocação teste	0.4107
s teste	0.0064

Sob qualquer intervalo de confiança corriqueiro (10%, 5% ou 1%) a hipótese nula (i.e. acurácia observada é igual à classificação aleatória) é rejeitada.

Os tempos de convergência foram medidos com o uso de um processador AMD Phenom II X2 550 3,10GHz, 4Gb RAM DDR2-1066 com 3,25Gb utilizáveis sob Windows 7 Ultimate 32 bits, programado sob Matlab (SCG da biblioteca NNtoolbox).

6.3 Sensibilidade aos Principais Parâmetros da MLP

A seguir a rede neural é submetida a testes de sensibilidade do desempenho em relação ao número de neurônios ocultos e ao critério de parada, que são dois principais parâmetros responsáveis por *under* ou *overfitting* da rede, mas que devem ser empiricamente testados para cada problema abordado com MLP.

A rede neural do teste anterior apresentou desempenho com confiança estatística excedente, mas com uma quantidade bastante elevada de parâmetros livres (301.203 sinapses). Isso pode ter um efeito dual sobre a resolução do problema, pelo *overfitting* de um lado, e pela complexidade dos padrões não-lineares apresentados à rede de outro. Como com qualquer rede neural aplicada, o teste sobre o número de neurônios ocultos é fundamental, já que não há de antemão um método explícito de definição.

Pela margem em ganho com eficiência em detrimento do desempenho, e pela possibilidade de treinar-se uma rede com menos parâmetros livres com menor risco de *overfitting*, o seguinte teste foi conduzido usando a mesma parametrização anterior, mas variando o número de neurônios ocultos entre 1 e 300. Desta forma chegou-se ao seguinte resultado:

Tabela 6.5 Sensibilidade ao número de neurônios ocultos

Neurônio Ocultos	300	100	30	10	7	5	3	2	1
melhor iteração	40.2	41.0	66.4	75.6	64.2	52.3	61.2	80.8	57.5
tempo (s)	46.88	17.37	12.90	10.15	8.80	6.63	8.49	10.63	5.26
mse treino	0.2091	0.2125	0.2120	0.2151	0.2254	0.2232	0.2590	0.2632	0.3261
mse validação	0.2198	0.2202	0.2198	0.2230	0.2331	0.2273	0.2688	0.2693	0.3319
mse teste	0.2204	0.2218	0.2213	0.2241	0.2347	0.2295	0.2659	0.2741	0.3399
s teste	0.0073	0.0140	0.0104	0.0118	0.0398	0.0284	0.0579	0.0625	0.0674
acurácia treino	0.4736	0.4602	0.4605	0.4442	0.4319	0.4131	0.3908	0.3932	0.3493
acurácia validação	0.4178	0.4070	0.4144	0.4045	0.3947	0.3909	0.3603	0.3750	0.3473
acurácia teste	0.4103	0.4079	0.4033	0.3944	0.3857	0.3722	0.3629	0.3581	0.3271
s teste	0.0066	0.0057	0.0058	0.0098	0.0116	0.0087	0.0072	0.0077	0.0079
z	3.17085	3.07282	2.88030	2.51344	2.15749	1.60177	1.21678	1.02064	-0.25790
p-valor	0.08%	0.11%	0.20%	0.60%	1.55%	5.46%	11.18%	15.37%	60.18%
-z	2.89900	2.83702	2.64038	2.10946	1.68048	1.24265	0.92142	0.70221	-0.58324
-z p-valor	0.19%	0.23%	0.41%	1.75%	4.64%	10.70%	17.84%	24.13%	72.01%
+z	3.44269	3.30862	3.12023	2.91742	2.63449	1.96089	1.51214	1.33907	0.06745
+z p-valor	0.03%	0.05%	0.09%	0.18%	0.42%	2.49%	6.52%	9.03%	47.31%
precisão treino	0.4730	0.4607	0.4585	0.4446	0.4357	0.4153	0.3991	0.3923	0.3522
precisão validação	0.4164	0.4059	0.4115	0.4055	0.3973	0.3931	0.3677	0.3744	0.3486
precisão teste	0.4090	0.4075	0.3995	0.3947	0.3878	0.3747	0.3703	0.3576	0.3331
s teste	0.0087	0.0070	0.0056	0.0077	0.0138	NaN	NaN	NaN	NaN
revocação treino	0.4736	0.4602	0.4606	0.4441	0.4319	0.4131	0.3906	0.3932	0.3491
revocação validação	0.4177	0.4076	0.4138	0.4051	0.3952	0.3909	0.3612	0.3750	0.3459
revocação teste	0.4107	0.4080	0.4036	0.3942	0.3852	0.3722	0.3630	0.3580	0.3291
s teste	0.0064	0.0064	0.0055	0.0096	0.0109	0.0049	0.0068	0.0076	0.0091

A primeira coluna equivale ao teste conduzido na seção 6.2. As linhas indicando “z” e “p-valor” se referem à estatística do teste de hipótese e ao nível descritivo da estatística para acurácia, e –z e +z faz referência ao intervalo de 1 desvio-padrão em cada direção.

As linhas com “s” indicam o desvio-padrão amostral. Repare que com 5, 3, 2 e 1 neurônio oculto existem ocorrências de *s* de teste com valor NaN, ou “Not a Number”. Isso acontece quando para alguma das 30 simulações a matriz de contingência tem uma coluna zerada em função de uma má classificação, mais provavelmente. Isso inviabiliza o cálculo da precisão média, que para esse caso fica com denominador igual a zero. Dessa forma, a variância amostral também é inviabilizada. Nenhuma solução alternativa foi sugerida para isso.

Um número muito reduzido de neurônios não traz resultados estatisticamente significativos. O ideal neste problema é usar ao menos 10. Entretanto, a melhora do desempenho da rede é exponencial à medida que o número de neurônios ocultos aumenta,

entre todos os indicadores medidos. Há uma grande vantagem na passagem de 1 para 10 neurônios, provavelmente em função da característica não linear do problema.

A Figura 6.6 apresenta dois gráficos plotados sobre o mesmo eixo de parâmetros. O primeiro mostra as métricas de desempenho (acurácia, precisão e revocação) em função do número de neurônios ocultos. E no segundo estão plotadas as linhas de nível descritivo da acurácia. A linha intermediária é o nível descritivo da classificação média dentre as 30 simulações rodadas. As duas linhas periféricas mostram o nível descritivo do desvio da acurácia média.

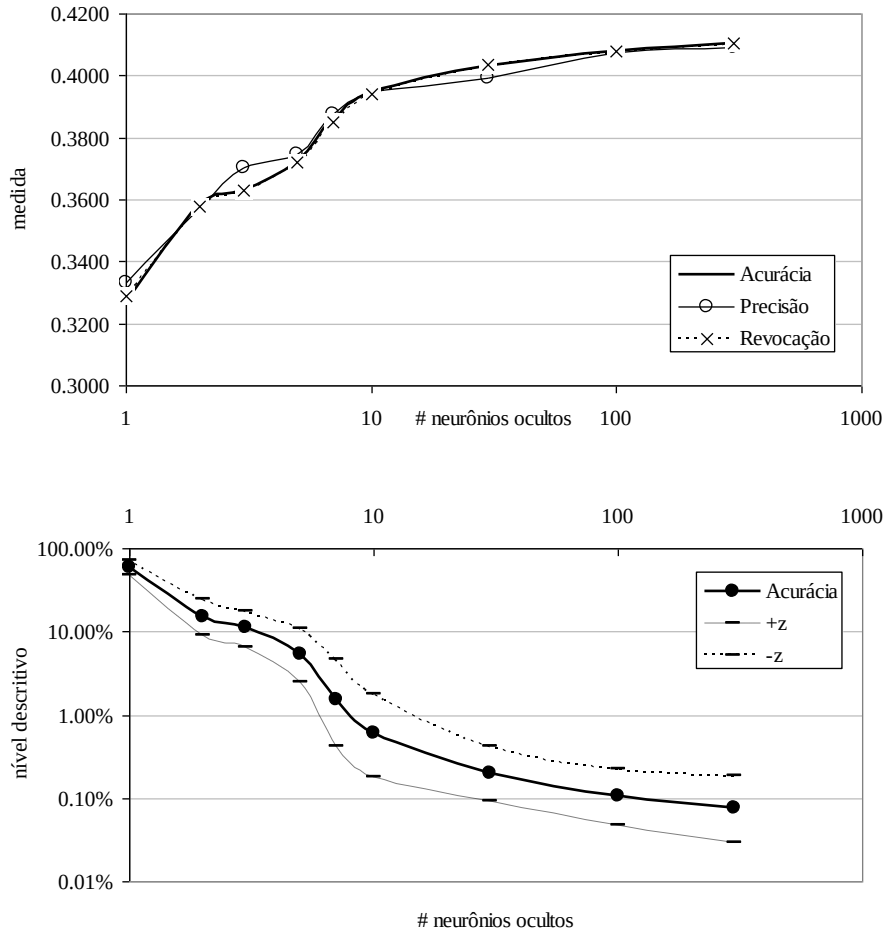

Figura 6.6 Número de neurônios ocultos: Curvas de desempenho e Nível descritivo da acurácia

O numero de iterações média para convergência (melhor iteração) é sensível ao número de neurônios ocultos, o que indica um mapeamento mais rápido do espaço não linear da dimensão léxica.

Mesmo com o número de iterações de convergência reduzido para redes mais complexas, o custo computacional quadrático entre as camadas de entrada e oculta (número de sinapses igual ao produto de entradas e neurônios ocultos) ainda favorece o tempo computacional de redes menores.

Com menos de 10 neurônios o número de iterações volta a decrescer. Isso ocorre pela convergência da rede a mínimos locais, evidenciado pelo aumento significativo do MSE.

A rede n-1-1 é o que mais se aproxima de uma regressão logística (n a dimensão de entrada), indicando que classificadores lineares podem não ser capazes de descrever bem o espaço de características. Por outro lado, a regressão logística converge para o mínimo global, e permite o teste de significância dos parâmetros, o que permite reduzir a dimensão do problema. Este é um problema a ser investigado em estudos posteriores.

O critério de parada adotado foi o de validação cruzada, que acompanha o desempenho da aprendizagem generalizada dos padrões com uma amostra disjunta àquela usada no treinamento. Se o desempenho (MSE) do conjunto de validação não melhorar numa nova iteração de treinamento, é adicionada uma contagem de falha. À medida que o número de falhas consecutivas aumenta, pode-se deduzir que a rede está se tornando mais propícia a *overfitting*. O processo é conduzido até que um número arbitrário e pré-estipulado de falhas seja atingido. A tabela a seguir mostra os resultados de mesma parametrização quando o número máximo de falhas no decremento do MSE de validação é alterado.

Tabela 6.6 Sensibilidade às falhas de validação

Max Falhas	128	64	32	16	8	4	2	1
melhor iteração	80.9	72.3	62.2	57.4	40.7	23.7	14.9	9.9
tempo (s)	17.14	14.30	11.15	10.97	8.29	5.06	3.23	2.54
mse treino	0.2049	0.2079	0.2163	0.2119	0.2216	0.2232	0.2322	0.2435
mse validação	0.2174	0.2163	0.2229	0.2192	0.2247	0.2249	0.2324	0.2416
mse teste	0.2198	0.2198	0.2245	0.2194	0.2267	0.2269	0.2331	0.2432
s teste	0.0028	0.0030	0.0131	0.0070	0.0156	0.0132	0.0290	0.0315
acurácia treino	0.4893	0.4729	0.4569	0.4573	0.4189	0.3985	0.3614	0.3452
acurácia validação	0.4162	0.4238	0.4141	0.4106	0.4001	0.3871	0.3629	0.3534
acurácia teste	0.4046	0.4022	0.4019	0.4070	0.3879	0.3716	0.3502	0.3375
s teste	0.0031	0.0038	0.0046	0.0052	0.0089	0.0040	0.0058	0.0086
z	2.93480	2.83673	2.82219	3.03283	2.24831	1.57636	0.69373	0.17071
p-valor	0.17%	0.23%	0.24%	0.12%	1.23%	5.75%	24.39%	43.22%
-z	2.80822	2.68140	2.63330	2.81675	1.88148	1.41292	0.45607	-0.18211
-z p-valor	0.25%	0.37%	0.42%	0.24%	3.00%	7.88%	32.42%	57.23%
+z	3.06137	2.99205	3.01108	3.24890	2.61513	1.73979	0.93138	0.52353
+z p-valor	0.11%	0.14%	0.13%	0.06%	0.45%	4.09%	17.58%	30.03%
precisão treino	0.4886	0.4724	0.4580	0.4563	0.4181	0.3967	0.3619	0.3457
precisão validação	0.4145	0.4218	0.4147	0.4089	0.3994	0.3844	0.3626	0.3539
precisão teste	0.4025	0.4005	0.4040	0.4043	0.3877	0.3687	0.3508	0.3373
s teste	0.0041	0.0039	0.0046	0.0049	0.0114	0.0082	0.0067	0.0097
revocação treino	0.4893	0.4728	0.4569	0.4573	0.4187	0.3986	0.3614	0.3452
revocação validação	0.4165	0.4236	0.4147	0.4114	0.4003	0.3865	0.3626	0.3534
revocação teste	0.4045	0.4027	0.4015	0.4064	0.3887	0.3717	0.3501	0.3374
s teste	0.0027	0.0034	0.0048	0.0051	0.0099	0.0054	0.0054	0.0093

Na Figura 6.7 de nível descritivo para validação cruzada é possível ver como uma tolerância a falhas elevada reduz o desempenho da inferência, por sujeitar e rede a treinamento excessivo. A partir de 16 falhas consecutivas o erro tipo I volta a aumentar para os 3 indicadores observados.

Com parada a 8 falhas consecutivas já se obtém um bom desempenho da MLP. Porém parar na primeira ou em poucas tentativas não é recomendado para esse tipo de problema, por não permitir à rede neural minimizar o erro suficientemente.

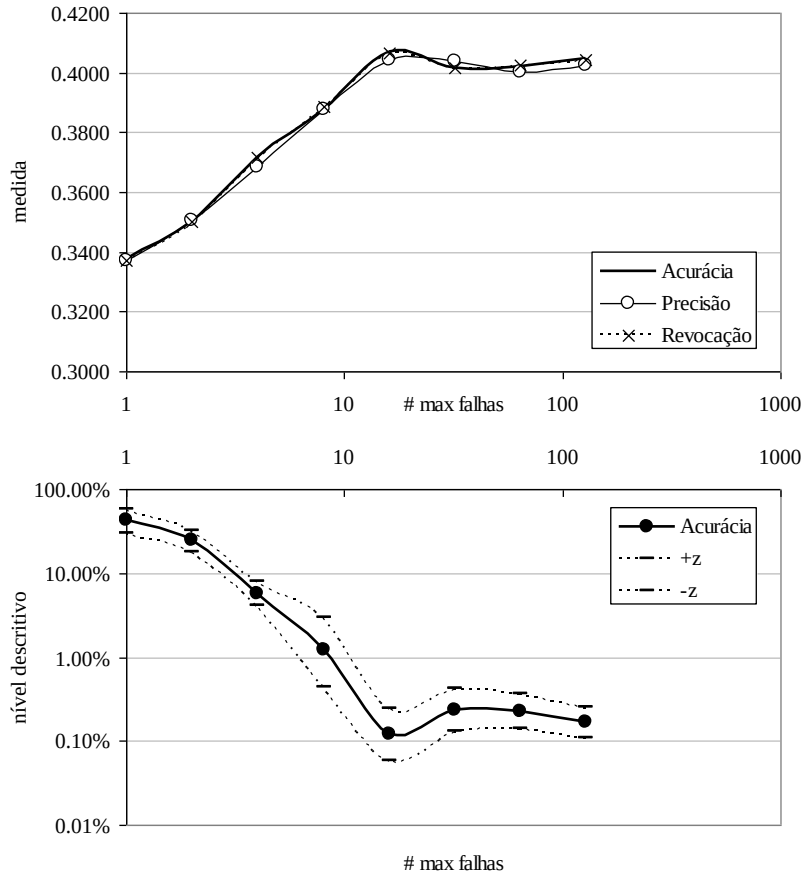

Figura 6.7 Número máximo de falhas: Curvas de desempenho e Nível descritivo da acurácia

O critério de validação cruzada se demonstrou robusto na seleção do nível de aprendizagem, e mesmo com o aumento de *overfitting* com alta tolerância a falhas, os testes mantiveram significância acima de 99%.

É fundamental notar que esse parâmetro de aprendizagem está muito relacionado com o algoritmo utilizado: SGC é rápido, e diversos algoritmos como o gradiente descendente, costumam convergir mais lentamente, levando à necessidade de um número máximo de falhas maior para evitar o risco de interromper a aprendizagem enquanto a rede ainda está aprendendo.

6.4 Regressão: Funciona?

O mesmo problema foi abordado com uma MLP de regressão não-linear, com função de ativação linear na camada de saída (identidade), e com o vetor de saída original, contendo os retornos do ativo-alvo relativos a cada observação. A Tabela 6.7 exibe a parametrização usada nas regressões seguintes.

Tabela 6.7 Parametrização da regressão

Parâmetros	
Filtro de documentos	TI stocks
Seleção de atributos por	T
Δt agrupamento	1min
Overnight	não
Pesagem	T
Poda de Atributos	1000
Normalização	n/a
Viés	1
y saída	S5INFT
y transformação	retorno líquido +20min
Classes	n/a
Método de inferência	mlp
Função de saída	identidade
Algoritmo	SCG
Neurônios ocultos	3 a 300
Amostra de treinamento	70% (1764)
Amostra de validação	15% (378)
Amostra de teste	15% (378)
Máximo falhas	30
Número de simulações	30

O desempenho da regressão variando o número de neurônios ocultos é mostrado na Tabela 6.8 a seguir. As taxas de precisão e revocação foram substituídas pelo coeficiente de correlação linear de Pearson da regressão (R^2). A acurácia é direcional – foi calculada com base na proporção de valores estimados cujo sinal equivale aos sinais realmente observados.

Tabela 6.8 Desempenho da regressão por número de neurônios ocultos

Neurônio Ocultos	300	100	30	10	3
melhor iteração	144.5	90.6	55.23	33.03	21.8
tempo (s)	215.5	41.51	11.35	4.8	3.57
mse treino	5.77E-06	6.09E-06	6.42E-06	6.54E-06	6.69E-06
mse validação	1.45E-05	8.70E-06	7.30E-06	6.53E-06	6.77E-06
mse teste	1.85E-05	9.64E-06	7.00E-06	6.72E-06	6.64E-06
acurácia treino	0.5879	0.5660	0.5326	0.5198	0.5178
acurácia validação	0.5119	0.5151	0.5193	0.5134	0.5186
acurácia teste	0.5109	0.5143	0.5144	0.5148	0.5086
z	0.423279	0.555312	0.559195	0.574728	0.333964
p-valor	33.60%	28.93%	28.80%	28.27%	36.92%
R^2 treino	31.3%	24.4%	12.3%	6.6%	2.9%
R^2 validação	0.9%	0.9%	2.4%	2.2%	0.5%
R^2 teste	0.5%	2.5%	0.0%	1.1%	-0.3%
t	0.089198	0.477156	0.006382	0.21525	0.06399
p-val	92.90%	63.35%	99.49%	82.97%	94.90%

O MSE da regressão é menor em função de y, que saiu do domínio $\{0,1\}$ passando a $[-0.0197;\ 0.0224]$. O MSE de treinamento caiu à medida que novos neurônios ocultos foram adicionados, mas isso causou *overfitting* da rede, evidenciado no MSE de validação na passagem de 100 para 300 neurônios, e ainda mais cedo no teste, na passagem de 10 para 30 neurônios.

Os valores de R^2 e acurácia ficaram consideravelmente baixos, sem qualquer relevância estatística. Além disso, não há qualquer tendência de crescimento ou decrescimento do R^2, permanecendo muito próximo de zero em qualquer caso, e probabilidade próxima de 1 da amostra coletada ter correlação nula. Esse resultado vai ao encontro com as descobertas de Schumaker e Chen, que obtiveram uma regressão significativa usando SVR.

Foi testada em seguida a influência de um número maior de categorias na tarefa de classificação por uma rede com 10 neurônios ocultos e máximo de 30 falhas na validação cruzada. A motivação por trás disso está em aproximar a tarefa de classificação que foi bem-sucedida à de regressão, e comparar os resultados à medida que o contradomínio (i.e. os retornos) é gradativamente particionado em subconjuntos menores. No limite, teríamos um classificador para o mesmo número de observações. Entretanto, conjecturo que apenas um classificador não-linear com observações correlacionadas seria capaz de aprender padrões neste extremo, já que teria de buscar por padrões vizinhos. Apesar disso, não é correto afirmar que à medida que o número de classes do problema aumenta, ele e assemelha cada vez mais à tarefa de regressão, já que regressão utiliza uma função de custo sobre um espaço continuo, enquanto classificação usa o espaço discreto.

Tabela 6.9 Desempenho da classificação por número de classes

Número de classes	2	3	4	5	6	7	8	9	10
melhor iteração	43.5	75.6	74.2	79.7	81.9	85.2	98.6	94.4	86.3
tempo (s)	4.25	10.15	7.88	8.49	9.23	9.95	10.96	11.88	11.49
mse treino	0.2413	0.2151	0.1856	0.1649	0.1402	0.1248	0.1077	0.1017	0.1011
mse validação	0.2493	0.2230	0.1916	0.1683	0.1433	0.1268	0.1101	0.1032	0.1021
mse teste	0.2513	0.2241	0.1923	0.1689	0.1439	0.1272	0.1103	0.1033	0.1018
s teste	0.0034	0.0118	0.0250	0.0322	0.0227	0.0188	0.0009	0.0159	0.0273
acurácia treino	0.5776	0.4442	0.3533	0.2794	0.2430	0.2138	0.1979	0.1593	0.1322
acurácia validação	0.5347	0.4045	0.3089	0.2441	0.2035	0.1810	0.1499	0.1267	0.1060
acurácia teste	0.5127	0.3944	0.2951	0.2330	0.1917	0.1642	0.1413	0.1226	0.1067
s teste	0.0114	0.0098	0.0121	0.0043	0.0057	0.0059	0.0030	0.0035	0.0064
z	0.49310	2.51344	2.02060	1.60094	1.30481	1.18415	0.95521	0.70827	0.43376
p-valor	31.10%	0.60%	2.17%	5.47%	9.60%	11.82%	16.97%	23.94%	33.22%
-z	0.05052	2.10946	1.47776	1.39177	1.00627	0.85921	0.77890	0.49129	0.01819
-z p-valor	47.99%	1.75%	6.97%	8.20%	15.71%	19.51%	21.80%	31.16%	49.27%
+z	0.93568	2.91742	2.56344	1.81010	1.60334	1.50908	1.13152	0.92525	0.84934
+z p-valor	17.47%	0.18%	0.52%	3.51%	5.44%	6.56%	12.89%	17.74%	19.78%
precisão treino	0.5777	0.4446	0.3533	0.2808	0.2407	0.2097	0.1994	0.1575	0.1359
precisão validação	0.5346	0.4055	0.3087	0.2409	0.2016	0.1784	0.1479	0.1250	0.1060
precisão teste	0.5129	0.3947	0.2954	0.2294	0.1906	0.1590	0.1375	0.1207	0.1084
s teste	0.0149	0.0077	0.0130	NaN	NaN	NaN	0.0042	NaN	NaN
revocação treino	0.5776	0.4441	0.3534	0.2799	0.2428	0.2138	0.1978	0.1591	0.1321
revocação validação	0.5345	0.4051	0.3087	0.2428	0.2033	0.1817	0.1495	0.1267	0.1066
revocação teste	0.5129	0.3942	0.2952	0.2324	0.1923	0.1635	0.1421	0.1232	0.1069
s teste	0.0120	0.0096	0.0123	0.0039	0.0055	0.0062	0.0036	0.0038	0.0073

Nota-se de imediato que o desempenho geral de classificação reduz à medida que se incrementa o número de classes. Em contrapartida, com 2 classes o desempenho é abruptamente pior. Lanço a hipótese (não testada aqui) de que isso se deve ao fato da rede não conseguir separar palavras espúrias com 2 classes, forçando-a a classificar junto com palavras de maior impacto e aumentando o desvio das estimativas. À medida que o número de classes aumenta, os indicadores de desempenho (acurácia, precisão e revocação) tornam-se mais rigorosos com a redução da área de alvo correta ("mosca"), i.e. a probabilidade de acertar a diagonal principal da matriz de confusão.

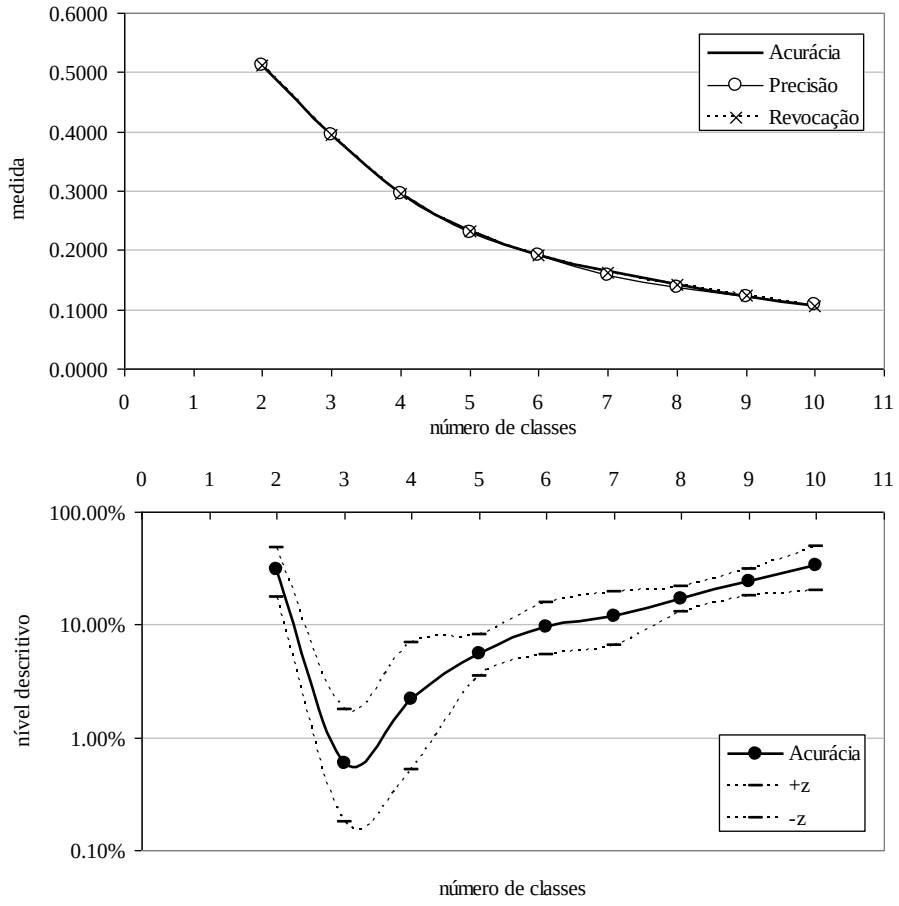

Figura 6.8 Número de classes: Curvas de desempenho e Nível descritivo da acurácia

Pelo exibido acima, a melhor opção com esta classe de problema é trabalhar com 3 classes equiprováveis. Com 3 classes ou mais, a rede é capaz de concentrar palavras espúrias nas classes intermediárias.

6.5 Efeito da Dimensão Léxica

A dimensão léxica define o ponto de corte aplicado sobre o número de palavras que a matriz de representação deve ter. Partindo de 0,56% da dimensão léxica total do corpus armazenada, o equivalente a 1.000 palavras, o número de atributos apresentados à rede neural foi sendo exponencialmente reduzido. Os resultados são mostrados a seguir.

Tabela 6.10 Desempenho da regressão por dimensão léxica

Dimensão	1000	512	256	128	64	32	16	8	4	2	1
melhor iteração	66.4	62.4	58.8	58.2	72.5	61.4	53.0	41.6	47.7	44.1	57.9
tempo (s)	12.90	8.09	5.72	5.16	5.64	4.58	3.83	1.92	2.32	2.10	2.83
mse treino	0.2120	0.2104	0.2122	0.2136	0.2178	0.2226	0.2267	0.2277	0.2280	0.2275	0.2260
mse validação	0.2198	0.2180	0.2176	0.2185	0.2214	0.2254	0.2282	0.2281	0.2289	0.2278	0.2280
mse teste	0.2213	0.2193	0.2193	0.2194	0.2235	0.2265	0.2288	0.2293	0.2297	0.2283	0.2285
s teste	0.0104	0.0032	0.0029	0.0025	0.0116	0.0149	0.0268	0.0153	0.0193	0.0178	0.0100
acurácia treino	0.4605	0.4596	0.4456	0.4371	0.4298	0.3940	0.3713	0.3614	0.3573	0.3450	0.3590
acurácia validação	0.4144	0.4158	0.4160	0.4005	0.4080	0.3711	0.3653	0.3560	0.3495	0.3304	0.3347
acurácia teste	0.4033	0.4106	0.3999	0.3986	0.3914	0.3608	0.3549	0.3393	0.3394	0.3295	0.3343
s teste	0.0058	0.0062	0.0136	0.0028	0.0052	0.0053	0.0041	0.0065	0.0044	0.0037	0.0093
z	2.88030	3.18176	2.74228	2.68779	2.38995	1.13325	0.88987	0.24699	0.25062	-0.15983	0.03994
p-valor	0.20%	0.07%	0.31%	0.36%	0.84%	12.86%	18.68%	40.25%	40.11%	56.35%	48.41%
-z	2.64038	2.92837	2.18348	2.57209	2.17446	0.91466	0.72128	-0.02028	0.06972	-0.31144	-0.34377
-z p-valor	0.41%	0.17%	1.45%	0.51%	1.48%	18.02%	23.54%	50.81%	47.22%	62.23%	63.45%
+z	3.12023	3.43516	3.30109	2.80349	2.60545	1.35184	1.05845	0.51427	0.43152	-0.00821	0.42365
+z p-valor	0.09%	0.03%	0.05%	0.25%	0.46%	8.82%	14.49%	30.35%	33.30%	50.33%	33.59%
precisão treino	0.4585	0.4579	0.4441	0.4350	0.4283	0.3936	0.3702	0.3606	0.3572	0.3456	0.3595
precisão validação	0.4115	0.4128	0.4142	0.3960	0.4054	0.3705	0.3639	0.3549	0.3487	0.3318	0.3350
precisão teste	0.3995	0.4081	0.3979	0.3953	0.3897	0.3602	0.3524	0.3377	0.3386	0.3299	0.3345
s teste	0.0056	0.0059	0.0114	0.0056	0.0051	0.0058	0.0109	NaN	0.0046	0.0082	0.0096
revocação treino	0.4606	0.4597	0.4456	0.4370	0.4298	0.3940	0.3713	0.3613	0.3574	0.3445	0.3591
revocação validação	0.4138	0.4153	0.4164	0.3998	0.4080	0.3712	0.3654	0.3561	0.3492	0.3314	0.3346
revocação teste	0.4036	0.4105	0.3996	0.3996	0.3917	0.3607	0.3543	0.3397	0.3391	0.3312	0.3341
s teste	0.0055	0.0063	0.0141	0.0029	0.0056	0.0052	0.0066	0.0039	0.0044	0.0053	0.0094

Aqui foi aplicada a poda entre uma e 1.000 palavras, como mostrado na Figura 6.9 em escala log-log.

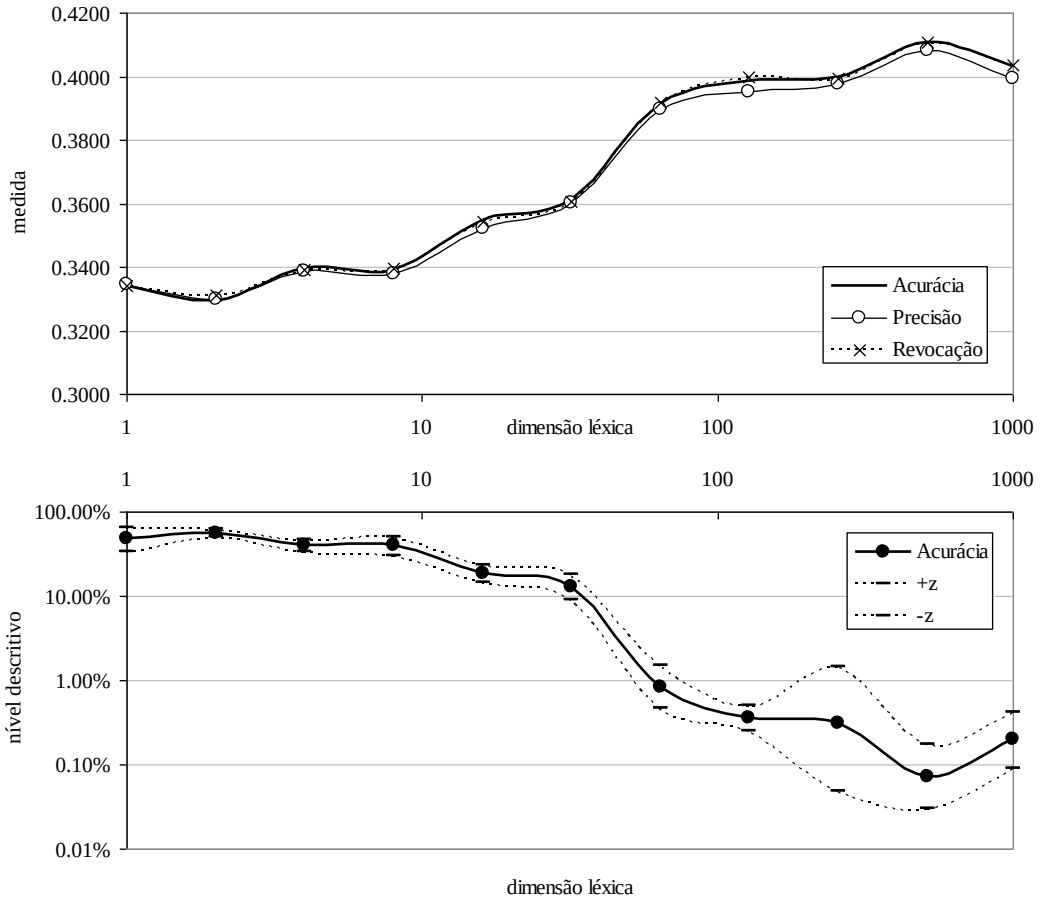

Figura 6.9 Dimensão léxica: Curvas de desempenho e Nível descritivo da acurácia

O desempenho sofre uma grande queda ao reduzir de 64 para 32 palavras-chave mais significativas, mantendo a confiança do teste satisfatória até 64 raízes de palavras. Desta forma, pode-se deduzir que as essas 32 palavras removidas são responsáveis pela grande diferença no desempenho da classificação. A Tabela 6.11 mostra os termos em questão.

Tabela 6.11 Termos removidos entre N=64 a N=32

our	43,050	will	32,372	system	26,180
blog	40,596	technolog	32,296	de	26,153
investor	40,331	busi	30,690	journalist	25,858
announc	39,448	be	30,029	media	25,748
site	38,305	includ	28,855	most	25,729
ar	37,519	solut	28,728	todai	25,364
contact	37,161	invest	27,989	manag	25,153
market	34,363	provid	27,962	call	24,928
center	34,026	ha	27,285	becom	24,465
all	33,332	their	26,815	sep	24,104
report	32,387	video	26,468		

Os termos em branco são majoritariamente ruído de informação: preposições, verbos básicos, e substantivos corriqueiros de um release, como *today* (*todai*), *blog* ou *contact*.

Em cinza claro estão palavras provavelmente mais impactantes para o setor de TI analisado, como *technology* (*technolog*), *solution* (*solut*), *system* e *provide* (*provid*), e para empresas de capital aberto em geral como *invest*, *market*, e *media*.

Em cinza escuro foram marcadas as palavras de (provavelmente) alto impacto, como *announcement* (*announc*), *report*, *september* (abreviado), *management* (*manag*) e *call*, que remetem a comunicados de relação com investidores.

Uma limitação metodológica nesse ponto do estudo é a impossibilidade de inferir quais sinapses da rede são mais relevantes na estimação, de forma que seja possível extrair conhecimento acerca de quais das 32 palavras foram mais determinantes na mudança de desempenho de classificação.

No presente estudo, são 64 entradas combinadas com 30 neurônios ocultos, além das sinapses da camada de saída, que adicionam a componente de não-linearidade das respostas. Isso significa que pulsos fracos de entrada podem gerar ativação de neurônios na saída, determinando a resposta da rede.

Para extração de conhecimento sobre os coeficientes da regressão, é necessário aplicar-se uma classificação linear tal como regressão logística.

6.6 Efeito da Pesagem

Nesta seção são testados 5 métodos de pesagem dos termos: Booleano (B), Contagem de Termos (T), Frequencia de Termos x Frequencia Inversa de Termos (TFIDF), Ganho de Informação (IG) e Informação Mútua (MI).

Tabela 6.12 Desempenho de classificação por Método de Pesagem

Pesagem	B	IG	MI	T	TFIDF
melhor iteração	19.8	57.5	60.7	66.4	48.6
tempo (s)	4.08	11.03	11.63	12.90	8.61
mse treino	0.2068	0.2145	0.2138	0.2120	0.2120
mse validação	0.2343	0.2220	0.2212	0.2198	0.2164
mse teste	0.2345	0.2238	0.2228	0.2213	0.2182
s teste	0.0117	0.0165	0.0111	0.0104	0.0032
acurácia treino	0.5075	0.4554	0.4612	0.4605	0.4543
acurácia validação	0.3963	0.4037	0.4090	0.4144	0.4255
acurácia teste	0.3896	0.3971	0.3999	0.4033	0.4082
s teste	0.0105	0.0083	0.0075	0.0058	0.0104
z	2.31730	2.62605	2.74228	2.88030	3.08369
p-valor	1.02%	0.43%	0.31%	0.20%	0.10%
-z	1.88366	2.28476	2.43197	2.64038	2.65492
-z p-valor	2.98%	1.12%	0.75%	0.41%	0.40%
+z	2.75093	2.96734	3.05260	3.12023	3.51247
+z p-valor	0.30%	0.15%	0.11%	0.09%	0.02%
precisão treino	0.5086	0.4551	0.4598	0.4585	0.4532
precisão validação	0.3950	0.4025	0.4062	0.4115	0.4236
precisão teste	0.3888	0.3956	0.3970	0.3995	0.4057
s teste	NaN	0.0075	0.0078	0.0056	0.0097
revocação treino	0.5075	0.4553	0.4612	0.4606	0.4543
revocação validação	0.3961	0.4037	0.4088	0.4138	0.4266
revocação teste	0.3896	0.3974	0.3999	0.4036	0.4074
s teste	0.0105	0.0073	0.0066	0.0055	0.0095

É curioso notar que apesar da pesagem booleana ter sido o pior método, teve um desempenho de treinamento muito superior aos demais métodos. Certamente houve *overfitting*, e seu uso correto deveria atentar a isso. Contudo o objetivo aqui é apenas mostrar as características de cada método sob as mesmas condições de teste.

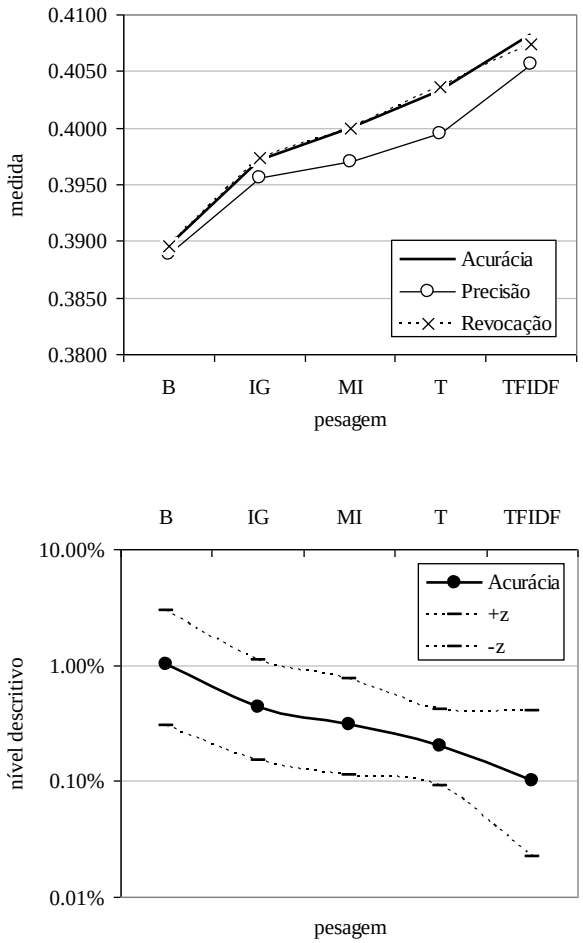

Figura 6.10 Métodos de Pesagem: Curvas de desempenho e Nível descritivo da acurácia

Dentre os métodos considerados, TFIDF e T tiveram os melhores resultados, o que mostra a capacidade de discriminação dos métodos mais simples. Porém, ao contrário da abordagem de Yang e Pedersen (1997), a dimensão léxica não foi testada em cruzamento com o método de pesagem, logo os resultados aqui apresentados são parciais.

Outra característica interessante foi a maior velocidade de convergência do TFIDF, quase tão rápida quanto a booleana.

Outros autores em FIR também estudaram a importância do método de pesagem usado. Peramunetillek (1997) comparou a representação de atributos com B, TFxIDF, TFxCDF e PCA (Principal Component Analysis) e observou um melhor desempenho com o terceiro.

Cho, Wütrich e Zang (1999) compararam T (normalizado entre 0 e 1) com outros 3 métodos de pesagem desenvolvidos no estudo e concluíram que o método tradicional apresenta desempenho tão bom quanto o melhor dos outros 3.

Dentre os estudos conduzidos no campo até hoje, é possível que o de Mittermayer e Knolmayer (2006) tenha sido o mais extenso. Os autores compararam 5 métodos de pesagem na seleção de atributos – aliás foi o único a estudar pesagem na seleção – e 4 métodos para representação. Os resultados indicam um melhor desempenho com CTF na seleção e WDFxIDF na seleção. Além disso, o número de atributos selecionados não alterou significativamente os resultados do teste.

6.7 Efeito do Tempo

O efeito das notícias sobre o tempo acerca de sua publicação é um tema largamente discutido em economia tratando de anúncios macroeconômicos e distribuição de resultados, contudo ainda não foi tocado pelos trabalhos em FIR. Inicialmente foi testada a capacidade de previsão dos movimentos do índice setorial de TI sobre intervalos de 5 minutos após a publicação, como mostrado na Tabela 6.13.

Tabela 6.13 Desempenho de classificação após a publicação

Instante	+5min	+10min	+15min	+20min	+25min	+30min
melhor iteração	56.8	51.7	69.9	66.1	63.6	55.8
tempo (s)	7.16	6.38	8.87	7.44	7.61	7.60
mse treino	0.2168	0.2163	0.2121	0.2113	0.2123	0.2149
mse validação	0.2226	0.2214	0.2210	0.2191	0.2198	0.2214
mse teste	0.2238	0.2221	0.2219	0.2204	0.2219	0.2240
s teste	0.0085	0.0024	0.0080	0.0034	0.0026	0.0072
acurácia treino	0.4178	0.4198	0.4550	0.4484	0.4443	0.4285
acurácia validação	0.3792	0.3879	0.3974	0.4024	0.4018	0.3897
acurácia teste	0.3698	0.3675	0.3924	0.3904	0.3795	0.3707
s teste	0.0071	0.0069	0.0141	0.0138	0.0043	0.0107
z	1.48346	1.38776	2.40002	2.31907	1.87733	1.51658
p-valor	6.90%	8.26%	0.82%	1.02%	3.02%	6.47%
-z	1.19659	1.10922	1.82649	1.75883	1.70080	1.08167
-z p-valor	11.57%	13.37%	3.39%	3.93%	4.45%	13.97%
+z	1.77033	1.66629	2.97356	2.87932	2.05386	1.95150
+z p-valor	3.83%	4.78%	0.15%	0.20%	2.00%	2.55%
precisão treino	0.4158	0.4189	0.4537	0.4468	0.4439	0.4273
precisão validação	0.3751	0.3874	0.3963	0.3998	0.4004	0.3876
precisão teste	0.3673	0.3671	0.3901	0.3865	0.3783	0.3694
s teste	0.0066	0.0075	NaN	0.0122	0.0048	0.0105
revocação treino	0.4178	0.4198	0.4548	0.4482	0.4443	0.4284
revocação validação	0.3783	0.3876	0.3977	0.4027	0.4020	0.3899
revocação teste	0.3703	0.3678	0.3928	0.3906	0.3798	0.3706
s teste	0.0073	0.0061	0.0129	0.0125	0.0042	0.0085

Os resultados mostram que a janela de 15 a 25 minutos após a publicação apresenta os melhores resultados de previsão. Contudo seria errôneo interpretar que o mercado leva alguns minutos para começar a incorporar informações textuais. O baixo nível descritivo dos testes logo após a publicação pode ser consequência de uma maior e imediata volatilidade sobre os retornos do ativo. Após 10 minutos, o desempenho da classificação aumenta em função da previsibilidade do movimento subsequente: é possível que a volatilidade reduza e o preço convirja a um patamar estável. Dessa forma, a janela de 15 a 25 minutos após o evento, representa o intervalo que o preço de mercado se ajusta. No entanto, neste trabalho não são realizados testes sobre a volatilidade, isso é deixado como sugestão para trabalhos futuros.

Depois de meia hora já não é possível prever o movimento com significância, o que indica que toda a informação já foi incorporada ao preço e que, possivelmente, outras informações já estão atuando sobre o movimento do ativo a partir dali.

Se existe atraso na incorporação da nova informação, mas o ajuste de fato acontece, então existe espaço para especulação sobre o preço do ativo, e possivelmente para auferir lucro dentro do intervalo de confiança observado. E se for este o caso, então não é possível corroborar a HEM.

A Tabela 6.14 complementa o estudo sobre a vizinhança temporal de publicação da notícia.

Tabela 6.14 Desempenho de classificação prévia à publicação

Instante	-30min	-25min	-20min	-15min	-10min	-5min
melhor iteração	59.7	55.3	46.0	49.6	50.9	45.8
tempo (s)	7.14	7.01	5.83	6.09	6.06	6.09
mse treino	0.2257	0.2187	0.2184	0.2210	0.2243	0.2192
mse validação	0.2308	0.2235	0.2225	0.2252	0.2288	0.2228
mse teste	0.2308	0.2243	0.2229	0.2265	0.2291	0.2237
s teste	0.0327	0.0096	0.0063	0.0126	0.0276	0.0083
acurácia treino	0.4100	0.4153	0.4097	0.4175	0.4077	0.3975
acurácia validação	0.3788	0.3839	0.3804	0.3920	0.3743	0.3734
acurácia teste	0.3649	0.3712	0.3659	0.3793	0.3719	0.3624
s teste	0.0078	0.0091	0.0124	0.0070	0.0068	0.0047
z	1.28470	1.53869	1.32147	1.86628	1.56811	1.18160
p-valor	9.94%	6.19%	9.32%	3.10%	5.84%	11.87%
-z	0.96601	1.17034	0.81723	1.58083	1.29165	0.99259
-z p-valor	16.70%	12.09%	20.69%	5.70%	9.82%	16.05%
+z	1.60338	1.90704	1.82572	2.15172	1.84457	1.37061
+z p-valor	5.44%	2.83%	3.39%	1.57%	3.25%	8.52%
precisão treino	0.4081	0.4184	0.4084	0.4178	0.4063	0.3952
precisão validação	0.3760	0.3855	0.3787	0.3918	0.3733	0.3709
precisão teste	0.3614	0.3736	0.3642	0.3799	0.3706	0.3591
s teste	0.0135	NaN	0.0097	0.0054	0.0143	0.0047
revocação treino	0.4098	0.4154	0.4098	0.4176	0.4075	0.3974
revocação validação	0.3789	0.3826	0.3796	0.3915	0.3749	0.3738
revocação teste	0.3655	0.3724	0.3661	0.3792	0.3720	0.3624
s teste	0.0094	0.0087	0.0088	0.0072	0.0072	0.0040

Antes da publicação os testes têm significância muito reduzida, mas a 15 minutos antes da divulgação existe alguma contração no erro do teste. A 5% o teste já é significativo, e talvez uma rede mais complexa seja capaz de categorizar esse movimento com maior precisão (ver seção 6.3 Sensibilidade aos Principais Parâmetros da MLP).

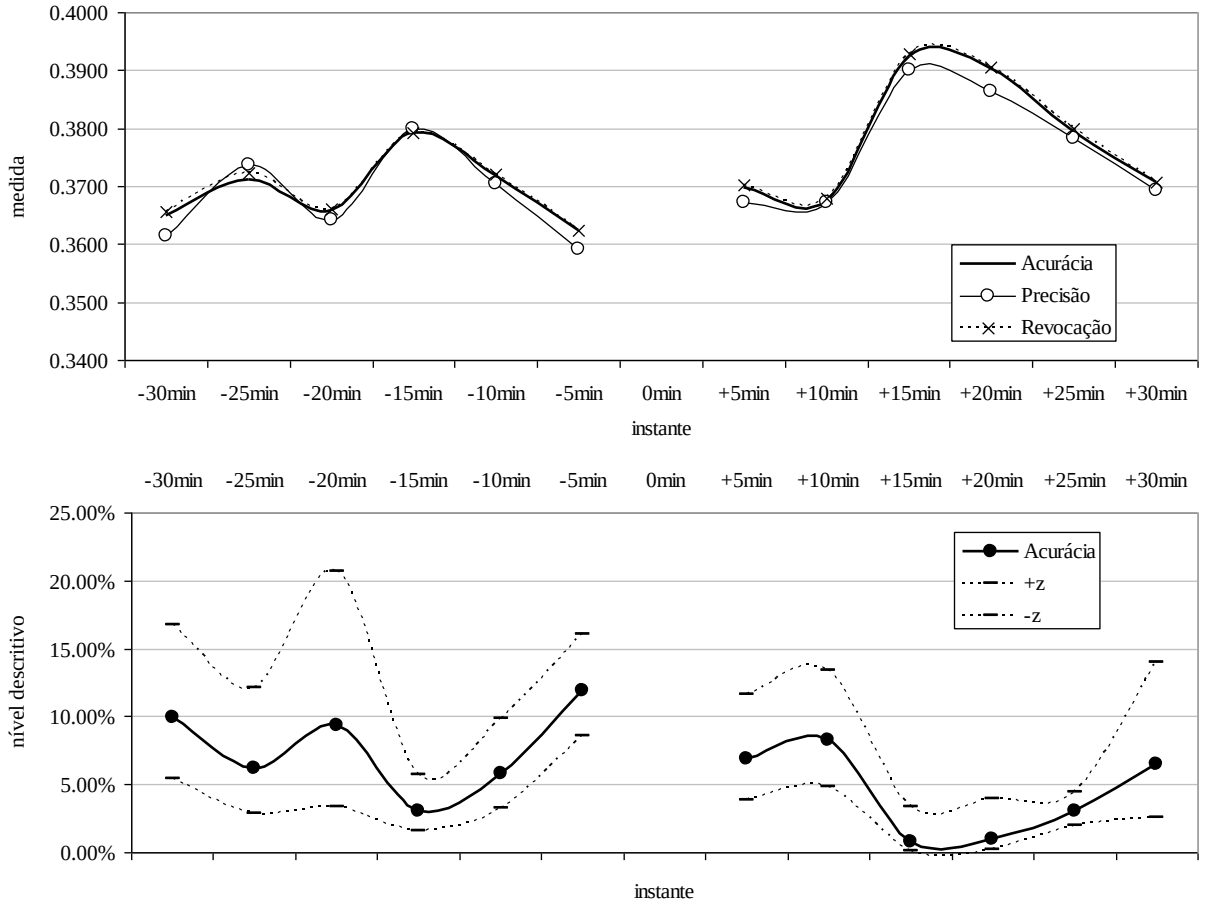

Figura 6.11 Tempo: Curvas de desempenho e Nível descritivo da acurácia

Com base na significância estatística dos resultados prévios à publicação da notícia, poder-se-ia concluir que há antecipação da informação a ser divulgada. Contudo, há uma diferença significativa entre os 15min e os demais momentos prévios. Se por ventura há qualquer vazamento de *insider information*, ele deve ocorrer por volta dos 15min anteriores ao fato publicado. Porém consideramos que a banda de variação do p-valor não se encontra completamente abaixo do nível de 5% para rejeitar a hipótese de antecipação.

CAPÍTULO 7 CONCLUSÃO

O trabalho executou uma série de experimentos com diversas parametrizações que envolvem a metodologia de recuperação de informação financeira. Primeiramente, os pressupostos assumidos para as afirmações de conclusão são listados. A seguir, as hipóteses para teste levantadas na seção 1.4 são examinadas, e as conclusões são apresentadas. Por fim, discorre-se sobre as contribuições do trabalho para a área de FIR e são feitas algumas sugestões para a continuidade do trabalho no campo.

7.1 Pressupostos assumidos

Os seguintes pressupostos são assumidos ao longo do trabalho:

- Hipótese da Recuperação de Palavras-Chave;
- A totalidade dos *press releases* é usada no período investigado;
- Perceptron Multicamadas com função de ativação tangente hiperbólica na camada de saída é representativo dentre qualquer classe de modelo de classificação;
- Perceptron Multicamadas com função de ativação linear na camada de saída é representativo dentre qualquer classe de modelo de regressão;
- O período temporal amostrado de cotações e notícias de 16/junho/2011 a 20/setembro/2011 é representativo sobre a população (i.e. sobre as séries temporais financeiras analisadas);

7.2 Hipóteses

As hipóteses definidas na seção 1.4 são revistas, e a conclusão e discussão sobre cada é dada a seguir:

$H_{0,1}$: O conteúdo textual de comunicados ao mercado não explica o movimento futuro de nenhum ativo financeiro.

Rejeita $H_{0,1}$. Foi visto logo no experimento 6.2 (Primeira Classificação) que o resultado de classificar o retorno do mercado de TI norte-americano em "Up", "Steady" e "Down" dentro de um horizonte de 20 minutos é significativamente positivo. O teste de hipótese foi realizado sobre a medida de acurácia observada sobre o conjunto de teste ($n = 378$), com média de 41,03% e desvio-padrão de 0,66% para 30 simulações. O intervalo de confiança para o valor descritivo ficou em $[0{,}03\%; 0{,}19\%]$.

Desta forma corroboram-se os mesmos achados prévios: Peramunetilleke (1997) obteve cerca de 52% de acurácia para o câmbio DEM/USD com uma amostra de teste de $n = 50$ e 3 classes equiprováveis, Schumaker e Chen obtiveram 71,18% de acurácia direcional para o retorno e amostra de tamanho não informado e Falinouss (2007) obteve 83% de acurácia com $n = 149$ para 2 classes quase equiprováveis (53%/47%). Todos os resultados com nível de confiança acima de 95%.

$H_{0,2}$: O mercado não antecipa a informação a ser publicada em uma notícia.

Não rejeita $H_{0,2}$. O resultado apresentado no experimento 6.7 (Efeito do Tempo) com $t = t_0 - 15min$ pode a princípio parecer dúbio com $\alpha = 5\%$ pela banda superior da curva de nível descritivo estar acima da probabilidade de erro aqui estabelecido. Em função disso, tomaremos a decisão de não rejeitar a hipótese. Apesar de α ser arbitrário, o nível descritivo é relativamente alto para afirmar que há antecipação da informação minutos antes de ser publicada.

A despeito da conclusão, esse é um resultado novo em FIR, que já foi abordado anteriormente nas pesquisas econômicas. Como exemplo, Patell e Wolfson (1979)

confirmaram a hipótese de que o investidor antecipa os efeitos da divulgação futura como reflexo sobre os preços de opções.

De qualquer forma, há uma diferença significativa entre os 15min que antecedem a publicação e os demais momentos prévios. Se há vazamento de *insider information*, ele deve ocorrer por volta dos 15min anteriores ao fato publicado.

7.3 Discussão Adicional

Foi observado no experimento 6.7 (Efeito do Tempo), que o mercado não incorpora imediatamente toda a informação publicada no preço dos ativos. De fato, foi previsto que o índice setorial de TI levou de 15 a 25 minutos após a publicação dos comunicados para ser ajustado, com nível descritivo inferior a 5%.

O fenômeno pode ser explicado como decorrência dos custos de transação ou do custo de captura da nova informação, ambos previstos por formas mais fracas da EMH. Dentre as duas possibilidades, a segunda é inadmissível, já que o custo de captura da informação é nulo (as informações são amplamente e gratuitamente divulgadas pelas agências de comunicação com o mercado). Existe ainda a possibilidade de ineficiência do processamento do conteúdo informacional pelos agentes, o que rejeitaria a EMH. Esta parece ser uma hipótese particularmente explicativa, já que os custos de transação tendem a zero à medida que o volume de especulação sobre um padrão recorrente no mercado aumenta.

A mesma hipótese de eficiência do mercado foi rejeitada e trabalhos anteriores, mas sob a ótica de ser possível auferir lucro operacional sobre o padrão detectado ou não. Wüthrich et. al. (1998) simulou um retorno médio de 30% a.a num período de 3 meses de previsão. Thomas e Sycara simularam um retorno de 20% num período de 200 dias. Schumaker e Chen (2009) obtiveram 8,5% de retorno simulado no período de 1 mês. Todos consideraram zero custo de transação e estratégias de investimento minimais em cima da previsão gerada (compra na previsão de alta, vende na de baixa).

7.4 Conclusões dos Experimentos

As seguintes conclusões são decorrentes da análise dos experimentos anteriores. É importante notar que elas se baseiam sobre certos pressupostos assumidos (ver seção 7.1) – e podem ser inválidas fora deles.

- Classificação preditiva de notícias pelo conteúdo léxico pode ser realizada com sucesso;
- A regressão preditiva de notícias é impossível de ser conduzida com sucesso: o R^2 é estatisticamente igual a zero;
- Quanto maior o número de neurônios ocultos da rede neural SCG com saída sigmoidal, melhor a classificação;
- Como critério de validação, 16 falhas de decremento do MSE de validação é o número ideal para o modelo neural considerado, sendo que há um ponto ótimo do parâmetro entre *under* e *overfitting*;
- A classificação com 3 classes equiprováveis tem maior significância estatística;
- A classificação com 2 classes equiprováveis é tão pobre quanto com muitas classes (10 ou mais);
- Pelo menos 128 palavras mais significativas são necessárias para ter-se confiança maior que 99% sobre os 3 indicadores considerados: acurácia, precisão e revocação, com o pré-processamento adotado;
- Com 1000 palavras o desempenho cai, mas o efeito de incorporar ruído com mais palavras não foi testado;
- A ordenação de desempenho de classificação por método de pesagem é: TFIDF > TF > MI > IG > B;
- O mercado ajusta o preço do ativo dentro de uma janela de 15 a 25 minutos após a publicação da notícia;
- O mercado não antecipa divulgação de informação privada;
- Os métodos e parâmetros classicamente empregados na literatura de FIR não são um intervalo estreito de ótimo local: diversos experimentos com parametrizações seccionalmente distintas tiveram confiança acima de 95%.

Dos experimentos realizados, 22 de 54 tiveram $\alpha < 5\%$. Isso considerando experimentos com elevada variação nos parâmetros testados: alguns inclusive com variação exponencial. A rigor esta é uma comparação que não prova nada, mas mostra a robustez da metodologia.

Na tarefa de classificação de notícias financeiras, vimos que o ponto de otimização global em termos de minimização do P-valor é difícil de alcançar por haver um encadeamento longo de processos dependentes, cada um com um amplo subespaço paramétrico.

Ademais, é desnecessário fazer o corte superior do dicionário ("podar a cabeça"), pois com mais de 30 palavras o modelo já começa a incluir palavras representativas.

7.5 Contribuições

Este trabalho contribuiu com:

- Validação do resultado positivo da metodologia de classificação empregada por autores anteriores (e.g., WÜTHRICH, PERMUNETILLEKE, *et al.*, 1998);
- Demonstração de algumas das dificuldades envolvidas no uso de regressão, como empregado por Schumaker e Chen (2009);
- Documentação da sensibilidade dos níveis de significância estatística da acurácia, precisão e revocação a diversos parâmetros do modelo;
- Rejeição da Hipótese do Mercado Eficiente;
- Não rejeição da hipótese de antecipação da informação.

7.6 Limitações do Trabalho

Apesar dos parâmetros que a metodologia implica terem sido extensamente testados, cada experimento fez uma seção transversal e testou parcialmente, ou independentemente, os parâmetros. Contudo, como é uma tarefa encadeada, é claro que os parâmetros são interdependentes, e com isso, apenas uma pequena amostra de regulagens foi apresentada, com centro na parametrização já empregada na literatura.

Portanto é possível que existam outras metodologias ou diferentes parametrizações que tenham resultados de classificação textual superior aos encontrados aqui. Os resultados já divergem bastante entre essa pesquisa e a de Schumaker e Chen (2009), no que diz respeito à capacidade de regressão dos dados, contudo, os autores usaram uma *Support Vector Regression*, cujas propriedades são distintas das redes neurais.

Outras limitações são o escopo de notícias restrito aos comunicados ao mercado, a falta do teste da influência de notícias publicadas fora do horário do pregão e uma grande quantidade de métodos de pesagem encontrados na literatura que não foram testados.

7.7 Sugestão de Trabalhos Futuros

A principal questão que ficou de ser investigada mais a fundo foi sobre o impacto das palavras individualmente sobre a classificação. Este trabalho ficou limitado ao uso de redes neurais que dificultam a extração de conhecimento dos parâmetros, e inviabilizam o teste estatístico sobre eles. Desta forma, não foi possível saber qual palavra-chave teve mais impacto e qual teve menos, o que seria um conhecimento valioso a este campo de pesquisa.

Uma possibilidade que fica aqui sugerida é de usar a regressão logística robusta. Se a regressão logística tiver sucesso na separação das classes seria um excelente indício para averiguar quanto a ocorrência de cada palavra individualmente impacta o valor futuro do ativo.

Outra coisa muito importante a se testar antes da iniciação de um trabalho como esse é a Hipótese de Recuperação de Palavras-Chave. Não foi encontrada nenhuma referência que discorresse sobre um teste preliminar sobre o corpus antes do restante da pesquisa ser desenvolvida. Talvez um teste Qui-Quadrado de independência sobre a matriz de freqüência pudesse ser formulado sob a interpretação dada em 1.4.

Um terceiro ponto a ser estudado no futuro é a volatilidade dos retornos no tempo. Em economia momentos de grau ainda maior – assimetria e curtose – já foram pesquisados. Desta forma questões como a apresentada em 6.7 (Efeito do Tempo) sobre o baixo nível descritivo em função da alta volatilidade logo após a publicação poderia ser testada. Além disso, uma elevada volatilidade também pode ser um sintoma de assimetria de informação nos instantes após a publicação da notícia.

Em último lugar, deixo uma ressalva sobre as abordagens na área de teste da lucratividade do modelo de previsão: ao passo que um experimento de *trading* simples e de custo transacional zero sobre a previsão é interessante para averiguar a EMH, qualquer mínima sofisticação num esquema de *trading* desse tipo corre o risco de perder sua validade científica. Dado que o conjunto de possíveis estratégias de trading sobre o modelo de previsão é infinito, certamente alguma trará lucro simulado acima do mercado. Para concluir o argumento, sugiro manter 3 diretrizes num estudo como esse: 1) testar a hipótese de retorno médio nulo com a aproximação dos retornos simulados por uma função de distribuição de probabilidade adequada; 2) Não utilizar ajustes paramétricos *a posteriori*; 3) Testar a estratégia de trading mais simples possível.

CAPÍTULO 8 BIBLIOGRAFIA

ACKLEY, D. H.; HINTON, G. E.; SEJNOWSKI, T. J. A Learning Algorithm for Boltzmann Machines. **COGNITIVE SCIENCE**, n. 9, p. 147-169, 1985.

AGGARWAL, R.; SCHIRM, D. C. Asymmetric impact of trade balance news. **Journal of International Financial Markets, Institutions and Money**, Cleveland, 8, 1998. 83-100.

ÄIJÖ, J. Impact of US and UK macroeconomic news announcements on the return distribution implied by FTSE-100 index options. **International Review of Financial Analysis**, Vaasa, 17, 2008. 242-258.

ALMEIDA, A.; GOODHART, C.; PAYNE, R. The Effects of Macroeconomic News on High Frequency Exchange Rate Behavior. **The Journal of Financial and Quantitative Analysis**, Vol. 33, n. No. 3, Setembro 1998. 383-408.

ALVARO ALMEIDA, C. G. R. P. The Effects of Macroeconomic News on High Frequency Exchange Rate Behavior. **The Journal of Financial and Quantitative Analysis**, v. Vol. 33, No. 3, p. 383-408, Sep. 1998.

APTÉ, C.; DAMERAU, F.; WEISS, S. M. Automated Learning of Decision Rules for Text Categorization. **ACM Transactions on Information Systems**, 1994. IBM Research Report RC 18879.

ARAMPATZIS, A. T. et al. **Phrase-based Information Retrieval**. Catholic University of Nijmegen. Nijmegen, p. 21. 1998.

ARNTZEN, T. C. et al. www.php.net. **PHP.net**, p. http://www.php.net/credits.php, 2011. Disponivel em: <http://www.php.net/credits.php>. Acesso em: 2011.

AROUCK, O. **Atributos de Qualidade da Informação**. Dissertação apresentada ao programa de Ciência da Informação da Faculdade de Ciência da Informação da UnB. Brasília, p. 117. 2011.

BASEL M. A. AWARTANI, V. C. Predicting the volatility of the S&P-500 stock index via GARCH models: the role of asymmetries. **International Journal of Forecasting**, v. 21, p. 167– 183, 2005.

BEALE, M. H.; HAGAN, M. T.; DEMUTH, H. B. **Neural Network Toolbox™ User's Guide - R2011b**. Natick, MA: The MathWorks, Inc., v. Version 7.0.2, 2011.

BEBER, A.; BRANDT, M. W. When It Cannot Get Better or Worse: The Asymmetric Impact of Good and Bad News on Bond Returns in Expansions and Recessions. **Review of Finance**, 14, 2010. 119-155.

BODIE, Z.; KANE, A.; MARCUS, A. J. **Fundamentos de Invetimento**. 3ª. ed. Porto Alegre: Bookman, 2000.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à Inferência Etatística**. 2. ed. Rio de Janeiro: Sociedade Brasileira de Matemática (Coleção Matemática Aplicada), 2010.

BOVESPA, B. **Manual da Política de Divulgação**. MANUAL DA POLÍTICA DE DIVULGAÇÃO DE ATO OU FATO RELEVANTE E NEGOCIAÇÃO DE VALORES MOBILIÁRIOS. [S.l.]: Anexo 1 da Ata da Reunião do Conselho de Administração da BM&F BOVESPA S.A. realizada em 08/05/08 –16h00m. 2008. p. 16.

BRÜCHER, H.; KNOLMAYER, G.; MITTERMAYER, M.-A. **Document Classification Methods for Organizing Explicit Knowledge**. University of Bern. Berna, p. 26. 2002. Em: Proceedings of the Third European Conference on Organizational Knowledge, Learning, and Capabilities. Athens, Greece: (2002).

CALVET, L. E.; FISHER, A. J. Multifrequency news and stock returns. **Journal of Financial Economics**, 86, 2007. 178-212.

CHO, V.; WÜTHRICH, B.; ZHANG, J. **Text Processing for Classification**. Computer Science Department - The Hong Kong University of Science and Technology. [S.l.], p. 30. 1999.

CHO, V.; WÜTHRICH, B.; ZHANG, J. Text Processing for Classification. **Journal of Computational Intelligence in Finance**.

DRAPER, N. R.; SMITH, H. **Applied Regression Analysis**. 3ª. ed. [S.l.]: John Wiley & Sons, 1998.

DUMAIS, S. et al. **Inductive Learning Algorithms and Representations for Text Categorization**. Microsoft Research & Stanford University. [S.l.], p. 8. 1998. Proceedings of the seventh international conference on Information and knowledge management.

FALINOUSS, P. **Stock Trend Prediction Using News Articles**. Lulea University of Technology - MSc thesis. [S.l.], p. 165. 2007.

FAMA, E. F. Efficient Capital Markets: A Review of Theory and Empirical Work. **The Journal of Finance**, 25, No.2, Maio 1970. 383-417.

FAMA, E. F. Efficient Capital Markets: II. **The Journal of Finance**, XLVI, No.5, Dezembro 1991. 1575-1617.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data**. Cambridge: Cambridge University Press, 2007.

FLORES, F. N. **Avaliando o Impacto da Qualidade de um Algoritmo de Stemming na Recuperação de Informações**. UFRGS. Porto Alegre, p. 48. 2009.

FUNG, G. H. I. C.; YU, J. X.; LAM, W. **Stock Prediction: Integrating Text Mining Approach using Real-Time News**. Department of Systems Engineering and Engineering Management - The Chinese University of Hong Kong. Hong Kong, p. 8. 2003.

FUNG, G. P. C.; YU, J. X.; LAM, W. **News Sensitive Stock Trend Prediction**. The Chinese University of Hong Kong. Shatin, p. 13. 2002. Copyright: Springer-Verlag.

GAJEWSKI, J.-F. Earnings announcements, asymmetric information, trades and quotes. **European Financial Management**, v. 5, n. 3, p. 411-423, 1999.

GERARD SALTON, M. J. M. **Introduction to Modern Information Retrieval**. NY: McGraw-Hill, Inc., 1986.

GIDÓFALVI, G. **Using News Articles to Predict Stock Price Movements**. University of California. San Diego, p. 9. 2001.

GITMAN, L. J. **Princípios de Administração Financeira**. 2ª. ed. Porto Alegre: Bookman, 2001.

GOODHART, C. A. E. et al. News Effects in a High Frequency Model of the Sterling Dollar Exchange Rate. **Journal of Applied Econometrics**, vol. 8, Jan-Mar 1993. 1-13.

GREENE, W. **Econometric Analysis**. 5th. ed. New Jersey: Prentice Hall, 2003.

GUJARATI, D. **Basic Econometrics**. 4th. ed. [S.l.]: The McGraw−Hill Companies, 2004.

HAYKIN, S. **Redes Neurais:** princípios e práticas. 2ª ed. ed. Porto Alegre: Bookman, 2001.

HURFORD, J.; HEASLEY, B. **Curso de Semântica**. Canoas: ULBRA, 2004.

INFORMATION_RETRIEVAL. Information_retrieval. **Information_retrieval**. Disponivel em: <http://en.wikipedia.org/wiki/Information_retrieval>. Acesso em: Novembro 2010.

JAMES M. PATELL, M. A. W. Anticipated information releases reflected in call option prices. **Journal of Accounting and Economics**, v. Volume 1, Issue 2, p. 117-140, 1979.

JAMES M. PATELL, M. A. W. The intraday speed of adjustment of stock prices to earnings and dividend announcements. **Journal of Financial Economics**, v. 13, n. 2, p. 223-252, June 1984.

JENNINGS, R.; STARKS, L. Earnings Announcement, Stock Price Adjustment, and the Exitence of Option Markets. **The Journal of Finance**, v. XLI, NO.1, March 1986.

JOHNSON, B. **Algorithmic Trading & DMA - An introduction to direct access trading strategies**. London: 4Myeloma Press, 2010.

KROHA, P.; BAEZA-YATES, R. **Classification of Stock Exchange News - Technical Report**. Universidad de Chile. [S.l.], p. 61. 2004.

LAVRENKO, V. et al. Language Models for Financial News Recommendation. **9th International Conference on Information and Knowledge Management**, Amherst, 2000. University of Massachusetts.

LIU, J. N. K.; DAI, H.; ZHOU, L. Intelligent Financial News Digest System. **Lecture Notes in Artificial Intelligence**, 2005. 112-120.

LOUHICHI, W. Adjustment of stock prices to earnings announcements: evidence from Euronext Paris. **Review of Accounting and Finance**, 7, No. 1, 2008. 102-115.

LOUHICHI, W. Adjustment of stock prices to earnings announcements: evidence from Euronext Paris. **Review of Accounting and Finance**, v. Vol. 7 No. 1, p. 102-115, 2008.

MA, Z.; SHENG, O. R. L.; PANT, G. Discovering company revenue relations from news: A network approach. **Decision Support Systems**, 47, 2009. 408-414.

MACHADO, C. M. V.; CORREIA, S. M.; CAPELLA, M. **Português 2:** Sintaxe. Rio de Janeiro: Senac Nacional, 2004.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **An Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2009. 581 p. DRAFT.

MARON, M. E. An Historical Note on the Origins of Probabilistic Indexing - Brief Communication. **Information Processing and Management**, n. 44, p. 971–972, 2007.

MCCORDUCK, P. **Machines Who Think**. 2ª ed. ed. Natick: A. K. Peters Ltd., 2004.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, vol. 5, 1943. 115-133.

MIRONOVA, S. Y. et al. **Advancements in Text Mining Algorithms and Software**. [S.l.]: [s.n.], 2004. 191-202; Cap. 22 p.

MITTERMAYER, M. A.; KNOLMAYER, G. F. **NewsCATS:** A News Categorization And Trading System. Proceedings of the Sixth International Conference on Data Mining (ICDM'06). [S.l.]: [s.n.]. 2006. p. 6.

MITTERMAYER, M. A.; KNOLMAYER, G. F. **Text Mining Systems for Market Response to News: A Survey**. University of Bern. Berna, p. 19. 2006.

MITTERMAYER, M.-A. Forecasting Intraday Stock Price Trends with Text Mining Techniques. **Proceedings of the 37th Hawaii International Conference on System Sciences**, 2004.

MOLLER, M. F. A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. **Neural Networks**, University of Aarhus, v. 6, p. 525-533, 10 January 1991.

MOLLER, M. F. A Scaled Conjugate Gradient Algorithm for Fast Supervised Learning. **Neural Networks**, v. 6, p. 525-533, 1993.

MÜLLER, A. L. et al. **Semântica Formal**. São Paulo: Contexto, 2003.

NG, A.; FU, A. W.-C. Mining Frequent Episodes for Relating Financial. **Lecture Notes in Artificial Intelligence**, Shatin, 2003. 27-39. Copyright: Springer-Verlag.

NOGUEIRA, B. M. **Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos**. Universidade de São Paulo. São Carlos, p. 104. 2009.

OLIVEIRA, P. Algoritmos de Radicalização - PT, 2010. Disponivel em: <http://student.dei.uc.pt/~pcoliv/reports/sigc_stemming.pdf>. Acesso em: 08 set. 2010.

PERAMUNETILLEKE, D. C. **A System for Exchange Rate Forecasting Using News Headlines**. Hong Kong University of Science and Technology. Hong Kong, p. 69. 1997.

PERAMUNETILLEKE, D.; WONG, R. K. **Currency Exchange Rate Forecasting from News Headlines**. University of New South Wales. Sydney, p. 9. 2000.

PORTER, M. F. An algorithm for suffix stripping. **Program**, 14, 1980. 130-137.

PRATYOOSH RAI, K. R. Comparison of Stock Prediction Using Different Neural Network Types. **International Journal of Advanced Engineering & Application**, v. 13, p. 4, Jan 2011.

ROSENBLATT, M. The Perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**, vol. 65, 1958. 386-408.

ROSS, S. **Probabilidade - Um curso moderno com aplicações**. 8ª. ed. [S.l.]: Bookman, 2010.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. **Learning internal representations by error propagation**. Cambridge: MIT Press, v. vol. 1, 1986. cap. 8.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations of back-propagation errors. **Nature**, Londres, vol. 323, 1986. 533-536.

SALTON, G.; BUCKLEY, C. Term Weighting Approaches in Automatic Text Retrieval. **Information Processing and Management**, 24, 1988. 513-523.

SCHUMAKER, R. P.; CHEN, H. A quantitative stock prediction system based on financial news. **Information Processing and Management**, 45, 2009. 571-583.

SEBASTIANI, F. **A Tutorial on Automated Text Categorisation**. Istituto di Elaborazione dell'Informazione - Consiglio Nazionale delle Ricerche. Pisa, p. 25. 1999.

SEO, Y.-W.; GIAMPAPA, J.; SYCARA, K. **Text Classification for Intelligent Portfolio Management**. Carnegie Mellon University. Pittsburgh, p. 20. 2002.

SEO, Y.-W.; GIAMPAPA, J.; SYCARA, K. **Financial News Analysis for Intelligent**. Carnegie Mellon University. Pittsburgh, p. 23. 2004.

SINGH, S. R.; MURTHY, H. A.; GONSALVES, T. A. Feature Selection for Text Classification Based on Gini Coefficient of Inequality. **JMLR: Workshop and Conference Proceedings 10: The Fourth Workshop on Feature Selection in Data Mining**, p. 76-85.

SOARES, M. V. B.; PRATI, R. C.; MONARD, M. C. Improvements on the Porter's Stemming Algorithm for Portuguese. **IEEE Latin America Transactions**, 7, No. 4, Agosto 2009.

SOLKA, J. L. Text Data Mining: Theory and Methods. **Statistics Surveys**, 2, 16 Julho 2008. 94-112.

SOUZA, E. C. O. D. **Criando Textos Segundo as Normas ABNT**. Scelisul. Vale do Ribeira, p. 41. 2005.

THOMAS, J. D. **News and Trading Rules**. Carnegie Mellon University. Pittsburgh, p. 214. 2003.

THOMAS, J. D.; SYCARA, K. **Integrating Genetic Algorithms and Text Learning for Financial Prediction**. Carnegie Mellon University. in Proceedings of the Genetic and Evolutionary Computing 2000 Conference Workshop on Data Mining with Evolutionary Algorithms, Las Vegas, p. 4. 2000.

VAN RIJSBERGEN, C. J. **Information Retrieval**. 2ª. ed. Londres: [s.n.], 1979. 147 p.

VIERA, A. F. G.; VIRGIL, J. Uma revisão dos algoritmos de radicalização em língua portuguesa. **informationr.net**, 2007. Disponivel em: <http://informationr.net/ir/12-3/paper315.html>. Acesso em: 08 nov. 2010.

WEI, C.-P.; DONG, Y.-X. A Mining-based Category Evolution Approach to Managing Online Document Categories. **Proceedings of the 34th Hawaii International Conference on System Sciences**, 2001.

WIKIPEDIA:/BUSINESS_WIRE. **Wikipedia**, 2011. Disponivel em: <http://en.wikipedia.org/wiki/Business_Wire>.

WÜTHRICH, B. et al. **Daily Prediction of Major Stock Indices from textual WWW Data**. The Hong Kong University of Science and Technology, The Chinese University of Hong Kong. Hong Kong, p. 6. 1998. Copyright: American Association for Artificial Intelligence.

WYSOCKI, P. D. **Cheap Talk on the Web: The Determinants of Posting on Stock Message Boards**. Univeridade de Michigan. [S.l.], p. 35. 1998.

YANG, Y.; CHUTE, C. G. An Example-Based Mapping Method for Text Categorization and Retrieval. **ACM TransactIons on Information Systems**, 12, No. 3, Julho 1994. 252-277.

YANG, Y.; PEDERSEN, J. O. **A Comparative Study on Feature Selection in Text Categorization**. [S.l.]. 1997.

YATES, R. B.; NETO, B. R. **Modern Information Retrieval**. [S.l.]: Adison Wesley Longman Limited, 1999.

ZHE, H.; CHIA-HAO, C. **Stemming Algorithms**. Recuperação de informação e tecnologias de recomendação: relatório intercalar. [S.l.], p. 30. N/D.

APÊNDICE A

Raízes mais Freqüentes

As 90 raízes de palavras (termos) mais freqüentes e suas respectivas contagens observadas no experimento da seção 6.2 (Primeira Classificação) são listados abaixo:

1	the	2157958	31	it	210580	61	free	121897
2	and	1808420	32	investor	209553	62	an	117836
3	of	1320834	33	blog	206156	63	jul	116391
4	to	1165172	34	ar	202094	64	your	113995
5	in	999433	35	site	199654	65	global	113021
6	a	933629	36	custom	199190	66	their	112996
7	for	876933	37	contact	189246	67	my	112595
8	new	563434	38	market	189201	68	provid	109896
9	pr	467488	39	de	175109	69	system	108867
10	on	429110	40	all	169455	70	onlin	105676
11	newswir	390174	41	report	169277	71	sep	104567
12	servic	357615	42	center	168038	72	research	104096
13	is	342093	43	be	167931	73	relat	104088
14	more	330440	44	announc	165483	74	packag	102730
15	us	329357	45	will	158595	75	manag	101385
16	at	309145	46	includ	157418	76	email	100203
17	with	299474	47	invest	148289	77	distribut	99513
18	et	298198	48	call	146434	78	send	99277
19	or	290435	49	ha	142381	79	search	99096
20	from	284265	50	journalist	139822	80	commun	97860
21	releas	281527	51	thi	136159	81	oper	97749
22	product	275341	52	million	134993	82	inc	97216
23	by	270381	53	financi	133638	83	these	96836
24	inform	246412	54	becom	132444	84	profit	95922
25	member	236804	55	most	130596	85	lead	95860
26	compani	235383	56	video	128140	86	feed	95455
27	as	233927	57	share	127598	87	todai	95410
28	about	230608	58	request	126226	88	newslett	94792
29	our	222266	59	media	125267	89	rss	93305
30	that	214321	60	busi	123771	90	brows	93211

APÊNDICE B

As tabelas seguintes são o resumo da revisão completa da literatura em FIR, desde o primeiro trabalho de 1997 até o este presente. Nela se apresenta um resumo da literatura completa, com descrição da metodologia, resultados e comentários. As primeiras colunas da tabela contêm metadados sobre os trabalhos (ano, autor, publicação, local) e as demais decompõem algumas características principais da metodologia.

Artigo	Ano	Autores	Publicação	Local	Objeto de Análise	Notícias	Período	Seleção	Representação	Inferência
A System for Exchange Rate Forecasting Using News Headlines	1997	Peramunetillek	Tese de Mestrado	Hong Kong	DEM/USD bid rates, JPY/USD bid rates "up", "down", "steady"	HFDF93 (Olsen) *dataset* : Manchetes de notícias do mercado monetário	01/out/92 a 30/set/93	Bag of words: 392 n-uplas de palavras-chave selecionadas por um especialista.	Usa 4 métodos de pesagem: BW, TFxIDF, TFxCDF, PCA	Geração iterativa de conjuntos de regras probabilísticas que minimizem o erro de estimação
Daily Prediction of Major Stock Indices from textual WWW Data	1998	Wüthrich, Permunetilleke, Leung, Cho, Zhang, Lam	Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining - 98	Hong Kong	Cotações de fechamento: DJIA, Nikkei225, FTSE 100, HSI, STI	The Wallstreet Journal (www.wsj.com): notícias dos últimos 100 pregões	N / D	Bag of words: >392 palavras-chave selecionadas por um especialista.	TFxCDF	Probalistic Rule Based, k-NN (k=9 com medida de similaridade euclideana), RNA (211 neuronios ocultos), e OLS
Cheap Talk On The Web: The Determinants of Posting on Stock Message Boards	1998	Wysocki	Working paper	EUA	Volume de postagens no Yahoo! message boards	Yahoo! Message Boards	dez/97 a ago/98	N/A	N/A	Estatística descritiva, OLS
Text Processing for Classification	1999	Cho, Wüthrich, Zhang	N/A	Hong Kong	Cotações de fechamento: Hang-Seng Index (HSI)	The Wallstreet Journal, Financial Times, CNN, International Herald Tribune e Bloomberg.	14/fev/97 a 6/nov/97	Bag of words: 392 n-uplas de palavras-chave	Teste de 4 métodos : Simple Weighting (SW), VWCR, Class-VWRD, e Cluster-VWRD.	Rule Based Machine. Gerador automático de populações de regras-base
Language Models for Financial News Recommendation	2000	Lavrenko, Schmill, Lawrie, Ogilvie, Jensen, Allan	Proceedings of the Internat.Conf. on Information and Knowledge Management,pages 389–396, 2000.	EUA	cotações de 10 minutos de 127 ações	38.469 notícias do Yahoo! Business com pelo menos 1 referência às ações estudadas	15/out/99 a 10/fev/00	N/A	N/A	Naive Bayes Classifier
Integrating Genetic Algorithms and Text Learning for Financial Prediction	2000	Thomas, Sycara	Proceedings of the Genetic and Evolutionary Computing	EUA	Cotações de fechamento e volume de negociação de 22 ações da Nasdaq e Nyse (Yahoo! Finance)	22 web bulletin boards mais populares em www.ragingbull.com	01/jan/99 a 31/dez/99	N / D	N / D	Naive Bayes Classifier, Algoritmos Genéticos e modelo híbrido
Currency Exchange Rate Forecasting from News Headlines	2001	Peramunetillek, Wong	N/A	Austrália	Taxas de Câmbio: USD-DEM, USD-JPY	HFDF93 (Olsen) *dataset* : Manchetes de notícias do mercado monetário	01/out/92 a 30/set/93	Bag of words: 392 n-uplas de palavras-chave selecionadas por um especialista.	BW, TFxIDF normalizado pelo máximo, TFxCDF	Geração de conjuntos de regras probabilísticas
Using News Articles to Predict Stock Price Movements	2001	Gidófalvi	N/A	EUA	12 ações NASDAQ	Igual a Lavrenko et.al. (2000) Cerca de7.000 notícias 5.500 notícias	14/nov/99 a 11/fev/00	MI: 1.000 atributos	N/A	Naive Bayes Classifier

Artigo	Ano	Autores	Publicação	Local	Objeto de Análise	Notícias	Período	Seleção	Representação	Inferência
News Sensitive Stock Trend Prediction / Stock Prediction: Integrating Text Mining Approach using Real-Time News	2002 / 2003	Fung, Yu, Lam	N/A	Hong Kong	614 ações da HKSX com média de 2.000 ticks/ação	Reuters Market 3000 Séries de 7 meses: 350.000 notícias	N / D	Clusters de segmentos de tendência dos preços são cruzados com clusters de documentos usando TFxIDF e TFxCDCxCSC	Comprimento dos vetores de documentos normalizados a |1|.	2 SVMs para "rise" e "drop"
Text Classification for Intelligent Portfolio Management / Financial News Analysis for Intelligent Portfolio Management	2002 / 2004	Seo, Giampapa, Sycara	N/A	EUA	Ações	1.239 / 6.239 notícias de diversos provedores: CNN, Forbes, Reuters, AP, PR Newswire, Business Wire e outros.	N / D	Bag of Words	TFxIDF normalizado	Naive Bayes Classifier + Expectation Maximization
News and Trading Rules	2003	Thomas	Tese de doutorado	EUA	Ações da Russel 3000 (em jul/01)	(1) Postagens de message boards no Yahoo! (2) Manchetes de notícias no Yahoo! Finance. 1682 ações em 1409 dias	(1) início de 98 a ago/01 (2) 01/mar/01 a 30/abr/02	N/A	N/A	AG, Regressão Linear (2) classificadores lógicos
Classification of Stock Exchange News Technical Report	2004	Kroha, Baeza-Yates	N/A	Chile	DAX30 e Nasdaq100	122.400 notícias Não indica a fonte	out/99 a set/03	1.000 palavras de maior ocorrência em cada classe	IDF	Naive Bayes Classifier
Forecasting Intraday Stock Price Trends with Text Mining Techniques	2004	Mittermayer	37th Hawaii International Conference on System Sciences	Suiça	Preços intraday e cotações bid/ask - Todas as ações no National Market System (NYSE, NASDAQ, AMEX, e 5 bolsas regionais)	PRnewswire filtrados (total de 6.602)	01/jan/02 a 31/dez/02	TFxIDF: 1.000 atributos	BW	SVM ("light": Thorsten Joachims, Cornell)
NewsCATS: A News Categorization And Trading System	2006	Mittermayer, Knolmayer	Proceedings of the Sixth International Conference on Data Mining (ICDM'06)	Suiça	Cotações das ações da S&P500 a 15 segundos	PRnewswire filtrados (total de 9.128)	01/abr/02 a 31/dez/02	CTFxIDF, CTF, CS, IG, OR: número de atributos entre 5% e 25% do número de documentos	WDFxIDF, BW, WDF, IDF	Rocchio, k-NN, SVM
Textual Analysis of Stock Market Prediction Using Financial News Articles / A quantitative stock prediction system based on financial news	2006 / 2009	Schumaker, Chen	N/A / Information Processing and Management	EUA	Ações, grupos setoriais de ações e S&P500	2.809 notícias do Yahoo! Finance e outros	26/out/05 a 28/nov/05	Nomes Próprios	BW	SVR
Stock Trend Prediction Using News Articles	2007	Falinouss	Tese de mestrado	Suécia	Ação da empresa Iran-Khodro (montadora de automóveis)	447 notícias financeiras e políticas sobre a companhia	2004 e 2005	CS: 200 atributos	TFxIDF	SVM com RBF
Information Retrieval from Press Releases	2011	Silva	Tese de mestrado	Brasil	Setor de TI GICS de ações	PRNewswire e Businesswire	16/jun/11 a 20/set/11	TF: 1 a 1.000 atributos	BW, TF, TFxIDF, IG, MI	ANN-SCG

Artigo	Metodologia	Resultados	Discussão dos Resultados
A System for Exchange Rate Forecasting Using News Headlines	Usa "keyword tuples", até quadrupla. Amostra de 50 observações para treinamento. Cerca de 50 observações para teste. Classes Up, Down e Steady equiprováveis.	TFxCDF foi em geral o melhor método de pesagem. O conjunto de dados DEM/USD produziu melhores resultados que JPY/USD. TFxCDF com DEM/USD teve 51%, 42% e 53% de acuracia em 1, 2 e 3h. Redes neurais tiveram 37% de acurácia. Aumentar o número de dados de treinamento não fez diferença significativa sobre o desempenho de previsão. As janelas de 1, 2 e 3 horas apresentam resultados médios de acurácia indistintos. Para DEM/USD, TFxCDF teve a melhor acurácia de classificação, com 51% numa janela de 1h. BW teve a melhor acuracia, com 48% na janela de 2h.	Introduziu a idéia de usar texto para prever mercados financeiros. O teste de hipótese sobre previsão aleatória não foi conduzido e a amostra testada foi relativamente pequena. Fica espaço para estudos futuros nessa direção.
Daily Prediction of Major Stock Indices from textual WWW Data	Não usa entradas numéricas. Treinamento sobre 100 dias: cotações de fechamento e união do conjunto de notícias no dia. Classes Up (>0.5%), Down(<-0.5%) ou Steady. Teste de previsão sobre 60 dias (dez/97 a mar/98) Usa informações disponíveis horas antes da abertura do pregão.	TFxCDF, com normalização pelo máximo, por ter sido o método de pesagem dos termos que resultou na melhor acurácia de previsão. Acurácia direcional média de: 43.7% com PRB, 44.4% com k-NN, 36.6% com RNA, 36.1% com OLS. Todos acima dos 33% de classificação aleatória. Retorno simulado sem custo de transação de 30% a.a. em 3 meses.	Não aponta os métodos de pesagem testados nem os resultados para chegar à conclusão de que TFxCDF é melhor.
Cheap Talk On The Web: The Determinants of Posting on Stock Message Boards	Usa postagens sobre 3.478 ações da NYSE e NADAQ para estatísticas descritivas, e 50 ações mais postadas para análise de série temporal (OLS)	Descobre que um maior volume acumulado de postagem sobre uma empresa é explicado por retornos passados extremos da ação, capitalização de mercado, relação P/E e M/B, volatilidade, volume de trades e cobertura de analistas. Há respectivamente 17.3%, 26.8% e 77.7% mais postagens na véspera, no dia e no dia seguinte do anúncio de resultados de uma empresa. Encontrou um coeficiente de 1,56% entre o volume de potagens no overnight e o volume operacional do dia seguinte. Coeficiente de 0.07% para retornos anormais do dia seguinte. A variação de postagens é autocorrelacionada, com coeficiente 0,23 entre um dia e o dia anterior.	Compara o volume de postagem sobre uma empresa com vários aspectos da ação e chega a diveros resultados interessantes utilizando apenas estatísticas simples. Ademais, a capacidade do volume de postagens prever volatilidade e retornos da ação no dia seguinte cria um elo entre pesquisas sobre economia da informação e classificação de notícias financeiras.
Text Processing for Classification	Mesma metodologia de Wüthrich et. al. (1998). Teste de diferentes formas de representação do documento. Teste de previsão sobre 79 dias.	Melhores métodos foram Simple Weighting (SW) e Vector Weighting with Cluster Relevance and Discrimination (Cluster-VWRD), com cerca de 43 a 46% de acurácia direcional.	Foca no processo de pré-processamento: 4 baseados em contagem de palavras-chave.
Language Models for Financial News Recommendation	Usa Modelos de Linguagem ao invés do clássico Modelo de Espaço Vetorial. 4 classes: "Surge", "Slight Rise", "Slight Fall" e "Plunge" As séries de preços são linearmente segmentadas por regressão e rotuladas por aglomeração de características	Compara o Modelo de Linguagem desenvolvido com um modelo de Espaço Vetorial e oberva que o segundo tem desempenho inferior bem como é incapaz de descriminar os padrões apresentados. Retorno simulado com 0.23% retorno médio por *trade* .	O método não permite controlar a resposta temporal a cada nova notícia (KROHA, 2004). Não é possível prever num dado instante usando um artigo que será publicado somente daqui a 20 minutos, mesmo que haja forte correlação com o movimento atual.
Integrating Genetic Algorithms and Text Learning for Financial Prediction	Duas metodologias de análise: (1) previsão baseada em classificação com máxima entropia para o texto, (2) algoritmo genético para aprender regras simples baseadas somente em dados numéricos de volume de trading, número de mensagens postadas e número de palavras postadas. Classes "up" e "down" para retornos diários. 52 dias de treinamento e 200 dias de teste.	As duas abordagens produzem retornos em excesso, mas a integração de ambas apresenta retornos excesivos (retorno estratégia - buy&hold > 0) superiores. Além disso, agregar múltiplos GAs numa única previsão aumenta o desempenho ainda mais. Quanto mais postagens sobre um ativo, melhor a classificação: Usando a metodologia (1), as 12 ações mais postadas tiveram um retorno excessivo de +6.91% e as outras 10 de -15.02%. Usando a metodologia (2), as 12 ações mais postadas tiveram um retorno excessivo de +5.94% e as outras 10 de -12.17%. Combinando ambas, o retorno excesivo foi de cerca de 20% no período de 200 dias.	O reultado em (1) não é suficiente para rejeitar a HME. Considera a revisão metodológica de (1) *ad hoc* , mas seria interesante testar se a abordagem baseada só em textos é suficiente para rejeitar a hipótese. Com 2 classes não há distinção de conteúdo textual ruidoso, e 52 dias de treinamento pode ser curto para uma série diária.
Currency Exchange Rate Forecasting from News Headlines	Igual a (PERAMUNETILLEK, 1997).	Hipótese nula de previsão aleatória foi testada, com um nível descritivo de 0,4% para acurácia de 51% com 3 classes, amostra de 60 e aproximação da distribuição binomial pela normal. Performance superior à Séries Temporais e Redes Neurais com input numérico.	Os resultados são os mesmos de (PERAMUNETILLEK, 1997), contudo a hipótese de previsão aleatória foi rejeitada.
Using News Articles to Predict Stock Price Movements	Desconsidera notícias fora do horário do pregão. Classes "up", "down" e "expected". Cerca de 4.300 notícias para treinamento e 1.300 para teste.	O poder preditivo do modelo proposto é baixo - previsor randômico apresentou acurácia superior, mas a correlação entre os movimentos de mercado e as notícias são altos.	Confirma que movimentos de curto prazo no preço de ações podem ser previsto a partir de artigos de notícias financeiras. Dentre os 20 min anteriores e posteriores ao lançamento do artigo, ele é capaz de prever o movimento. Contraria a EMH.

Artigo	Metodologia	Resultados	Discussão dos Resultados
News Sensitive Stock Trend Prediction / Stock Prediction: Integrating Text Mining Approach using Real-Time News	Série de preços segmentada. Os segmentos são aglomerados em 3 classes "rise", "drop" e "teady" de acordo com a inclinação da reta e com o R^2. A aglomeração é feita com um algoritmo propoto "Guided Clustering". 90% observações para treino e 10% para teste.	Classificação com acurácia significativamente superior à randômica.	Os números não são divulgados, de forma que não é possível avaliar quanto questões como benchmark e cenário econômico poderiam anular os retornos positivos do modelo.
Text Classification for Intelligent Portfolio Management / Financial News Analysis for Intelligent Portfolio Management	5 classes ('good', 'good uncertain', 'neutral', 'bad uncertain' e 'bad'). As notícias foram rotuladas manualmente. Adota o modelo de espaço vetorial [Salton, 1989] para análise textual.		Não explica a co-localização com detalhe: 2 palavras são co-localizadas dada uma distância de vizinhança ou dada a continência numa mesma frase?
News and Trading Rules	Realiza 2 estudos: (1) sobre postagens em fóruns de discussão no Yahoo! como feito em Wysocki (98) e (2) sobre manchetes de notícias. Em (2) os exemplos de treinamento são classificados à mão. As 40 classes de notícias são definidas de acordo com o tipo de notícia, e o objetivo da classificação não é previsão, mas identificação do tipo. Usa uma vasta combinação de índices técnicos como preditores de mercado.	(1) Endossa os resultados descritivos encontrados por Wysocki (98): acumulação exponencial de postagens por ações, correlação entre volume de postagem no overnight e volume de operação sobre uma ação no dia seguinte, correlação com o retorno anormal absoluto no dia seguinte. (2) Precisão de cerca de 90% para a maioria das classes de notícias, revocação entre 10% a 100%, dependendo da classe. Em particular, simulação de trading sobre notícias de divulgação de resultados tiveram retorno consistentemente positivo. Muitas outras classes não produziram resultados relevantes.	O achado é explicado através da tese de sobre-reação de De Bondt e Thaler (1985, 1987), que afirmam que seguindo grandes quedas de preço, o preço "ricocheteia" em períodos subsequentes.
Classification of Stock Exchange News Technical Report	Dividiu a série de cotações da DAX manualmente em 4 segmentos => 4 classes (Down2000, Down2002, Up1999 e Up2003). Metade dos documentos para treino e outra para teste.	Acurácia média de descrição de 90%, conclui que cada período de tempo tem uma linguagem particular pela qual pode ser identificado. Aglomerar as 4 classes em 2 reduziu a acurácia para cerca de 70%. Ressalva que em alguns textos a classificação não é fácil nem por frequencia nem pelos experts. Argumenta que um sistema de recomendação de notícias como em [Lavenkro, 2000] por vezes não é suficiente, pois podem sucitar opiniões divergentes dos experts.	Enumera exemplos de ambiguidade no uso do conjunto léxico para interpretação de textos. Mostra que períodos de tempo de longo prazo podem ser muito bem identificados por um conjunto característico de ocorrência de palavras. Um ponto a ser investigado é se a segmentação automática produziria resultados semelhantes.
Forecasting Intraday Stock Price Trends with Text Mining Techniques	Faz uma filtragem da base de notícias para remover ruído e redundância. Classes "Good News", "Bad News" e "No Movers". O critério de classificação é diferente do comum na literatura.	Revocação média de 58%, precisão média inferior a 33%. Retorno simulado médio de 0,20% por trade contra 0,07% de um trader aleatório.	Concebeu uma estratégia de trading que apesar de baixa precisão na classificação foi capaz de ter retornos consistentemente positivos.
NewsCATS: A News Categorization And Trading System	3 Classes: "good", "bad", "neutral" e uma 4a classe "unclear", para notícias incertas que são excluídas do treinamento. Foram filtradas categorias de comunicados pré-definidos pelo provedor com maior potencial de impactar o mercado. Diversos estudos de sensibilidade foram conduzidos: sobre seleção de atributos, dimensão do conjunto léxico, representação do documento e algoritmo de inferência.	CTF para seleção de atributos, WDFxIDF para representação do documento e SVM para inferência. Qualquer dimensão léxica entre 5% e 25% do número de observações.	Testou uma ampla gama de métodos de pesagem comummente encontrado em IR. Os autores não deixam claro qual a significância estatística de uma acurácia reportada de 83% para classes não equiprováveis: 81% das observações de treino e teste pertenciam à classe "neutral", o que certamente enviesa o resultado do classificador.
Textual Analysis of Stock Market Prediction Using Financial News Articles / A quantitative stock prediction system based on financial news	Filtragem das notícias: após abertura, antes do fechamento do leilão. Aplica regressão sobre a representação vetorial das notícias. Realiza treinamento por setor, indústria, ação e universal. 8.5% retorno simulado no período.	71% de acurácia direcional (o equivalente a 2 classes equiprováveis) para previsão por setor Retorno positivo e acima do benchmark, dos profissionais "experts" e da média de fundos quantitativos.	Primeiro trabalho em FNC a aplicar regressão sobre o conteúdo textual, e com sucesso. Propõe o uso de previsões para setores: mostra a melhor acurácia direcional e maior retorno.
Stock Trend Prediction Using News Articles	Segmentação da série de preços 2 classes: "rise" e "drop", 717 e 799 documentos em cada. 70% dos documentos para treino e 30% teste. 298 notícias de treino, 149 de teste.	83% de acurácia, 80% de revocação média e 84% de precisão média com 2 classes quase equiprováveis na amostra de 149 notícias. Erro tipo 1 bem abaixo de 1%.	
Information Retrieval from Press Releases	Aplica testes de sensibilidade sobre a metodologia já "consolidada" no campo. Testa o impacto das notícias ao longo do tempo no intraday.	Melhores parâmetros: TFxIDF, 3 classes, pelo menos 128 atributos. Classificação com significância estatística para acurácia, porém nula para regressão. A parametrização empregada na metodologia de FNC não é um ótimo local. O mercado ajuta o preço do ativo à notícia numa janela de 15 a 25 minutos após a publicação. Não há antecipação de informação.	O resultado com a regressão diverge daquele apontado por Schumaker (2007), contudo não deveria. Fica a questão para investigação futura.