



UMA METODOLOGIA PARA A PREVISÃO DO ÍNDICE BOVESPA
UTILIZANDO MINERAÇÃO DE TEXTOS

Elisangela Lopes de Faria

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Civil, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Civil.

Orientador(es): Nelson Francisco Favilla Ebecken
Marcelo Portes de Albuquerque

Rio de Janeiro

Julho de 2012

UMA METODOLOGIA PARA A PREVISÃO DO ÍNDICE BOVESPA
UTILIZANDO MINERAÇÃO DE TEXTOS

Elisangela Lopes de Faria

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA CIVIL.

Examinada por:

Prof. Nelson Francisco Favilla Ebecken, D.Sc.

Prof. Marcelo Portes de Albuquerque, D.Sc.

Profa. Beatriz de Souza Leite Pires de Lima, D.Sc.

Profa. Heloisa Márcia Pires, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

JULHO DE 2012

Faria, Elisângela Lopes de

Uma metodologia para previsão do índice BOVESPA utilizando Mineração de Textos/ Elisângela Lopes de Faria. – Rio de Janeiro: UFRJ/COPPE, 2012.

XI, 105 p.: il.; 29,7 cm.

Orientador: Nelson Francisco Favilla Ebecken

Marcelo Portes de Albuquerque

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia Civil, 2012.

Referências Bibliográficas: p. 94-103.

1. Previsão de séries temporais. 2. Mineração de Textos . 3. Mercados Financeiros. I. Ebecken, Nelson Francisco Favilla *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Civil. III. Título.

À minha mãe Maria e ao meu pai
José Lopes (*in memoriam*).

AGRADECIMENTOS

A Deus por ter me iluminado e me dado força e inspiração para conseguir concluir este trabalho.

Em especial aos meus pais, que apesar de nunca terem frequentado uma escola sempre me orientaram e me incentivaram à vida acadêmica. Agradeço o amor, apoio, carinho e dedicação de toda uma vida.

Ao meu marido, Jorge Luis pelo carinho, paciência e apoio ao longo do desenvolvimento deste trabalho. Suas palavras de incentivo foram fundamentais para a conclusão do mesmo.

À minha irmã Marlene, minhas sobrinhas Jane e Cyntia e meu cunhado Ronaldo pelo carinho e torcida em todos os momentos.

Ao meu orientador, prof. Dr. Nelson Francisco Favilla Ebecken, pelo grande apoio e confiança durante todo o desenvolvimento deste trabalho.

Ao meu co-orientador, prof. Dr. Marcelo Portes de Albuquerque, pelos ensinamentos transmitidos. Os anos sob sua supervisão no Centro Brasileiro de Pesquisas Físicas têm contribuído muito para meu crescimento intelectual.

Ao Prof. Dr. Márcio Portes Albuquerque do CBPF, pela contribuição dada neste trabalho.

À profa. Dra. Valéria Bastos por semear os primeiros conhecimentos sobre a área de mineração de textos.

A José Thadeu Cavalcante, engenheiro do CBPF pelas discussões sobre redes neurais e linguagens de programação.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

UMA METODOLOGIA PARA A PREVISÃO DO ÍNDICE BOVESPA
UTILIZANDO MINERAÇÃO DE TEXTOS

Elisangela Lopes de Faria

Julho/2012

Orientadores: Nelson Francisco Favilla Ebecken

Marcelo Portes de Albuquerque

Programa: Engenharia Civil

Nesta dissertação a técnica de mineração de textos é utilizada para prever o mercado de ações no Brasil, em particular o índice Bovespa, que representa o principal indicador do mercado acionário brasileiro. Para esta finalidade foram processadas notícias macroeconômicas e financeiras, divulgadas nos principais sites de notícias do Brasil, juntamente com as séries temporais das cotações do indicador estudado no período de fevereiro de 2010 a junho de 2011. Diferentes modelos de previsão foram testados com várias configurações de parâmetros. Os resultados obtidos foram quantificados através das métricas de avaliação acurácia, precisão, *recall* e medida-F e os mesmos demonstraram que é possível prever o Ibovespa com uma acurácia de 66%. Por fim uma estratégia de negociação foi desenvolvida e a mesma reportou uma lucratividade de 33%.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

A METHODOLOGY FOR BOVESPA INDEX FORECASTING
USING TEXT MINING

Elisangela Lopes de Faria

July/2012

Advisors: Nelson Francisco Favilla Ebecken
Marcelo Portes de Albuquerque

Department: Civil Engineering

In this thesis the technique of text mining is used to predict the stock market in Brazil, particularly the Bovespa index, which represents the main indicator of the Brazilian stock market. For this purpose were processed macroeconomic and financial news, published in major news sites in Brazil, along with time series of quotations from the indicator studied from February 2010 to June 2011. Different predictive models were tested with various parameter settings. The results were quantified using the metrics of evaluation accuracy, precision, recall and F-measure and they demonstrated that one can predict the Ibovespa with an accuracy of 66%. Finally a trading strategy was developed and it reported a profit of 33%.

Sumário

| | |
|--|----|
| 1. Introdução..... | 1 |
| 1.1 Objetivos..... | 3 |
| 1.2 Descrição e organização do trabalho | 5 |
| 2. Revisão da Literatura – Principais Trabalhos | 7 |
| 2.1 Arquitetura geral dos modelos de previsão do mercado e mineração de textos 7 | |
| 2.2 Resumos dos principais trabalhos..... | 9 |
| 3. Mercado de Ações no Brasil | 21 |
| 3.1 Características do Mercado de Ações | 22 |
| 3.2 Análises de Investimento | 25 |
| 3.2.1 Hipótese do Mercado Eficiente | 25 |
| 3.2.2 <i>Random Walk</i> | 26 |
| 3.2.3 Análise Fundamentalista | 26 |
| 3.2.4 Análise Técnica | 26 |
| 3.3 Análise Técnica x Análise Fundamentalista..... | 27 |
| 4. Mineração de Textos para Análise de Notícias..... | 28 |
| 4.1. Classificação de Textos | 29 |
| 4.1.1 Aplicações da Classificação de Textos..... | 29 |
| 4.1.2 Processo Geral de Classificação Automática de Textos..... | 30 |
| 4.1.2.1 Obtenção de Documentos | 32 |
| 4.1.2.2 Pré-processamento de Notícias..... | 34 |
| 4.1.2.3 Extração de Conhecimento | 43 |
| 4.1.2.4 Métricas de Avaliação | 47 |
| 5. Descrição do Sistema de Previsão do Mercado | 50 |
| 5.1 Coleta de Dados | 52 |
| 5.1.1 Base de dados de Séries Temporais Financeiras | 52 |
| 5.1.2 Base de dados de Notícias Financeiras..... | 54 |
| 5.1.3 Atribuição de classes aos Títulos das Notícias..... | 55 |
| 5.2 Pré-processamento das Notícias | 59 |
| 5.2.1 Pré-processamento no Exp1 | 60 |
| 5.2.2 Pré-processamento no Exp2..... | 61 |

| | | |
|-------|---|-----|
| 5.3 | Classificação dos documentos | 63 |
| 5.3.1 | Classificação de documentos no Exp1 | 64 |
| 5.3.2 | Classificação de documentos no Exp2 | 65 |
| 5.4 | Avaliação dos Sistemas | 66 |
| 5.5 | Estratégia de Negociação | 67 |
| 6. | Resultados e Análises | 70 |
| 6.1 | Base de Dados: Séries Temporais Financeiras e Notícias | 70 |
| 6.2 | Experimento 1. Resultados e Análise | 73 |
| 6.2.1 | Resultados obtidos Experimento1 | 73 |
| 6.2.2 | Análise e avaliação do Modelo no Exp1 | 75 |
| 6.3 | Experimento 2. Resultados e Análise | 77 |
| 6.3.1 | Resultados obtidos Experimento2 | 77 |
| 6.3.2 | Análise e avaliação do Modelo no Exp2 | 82 |
| 6.4 | Comparação entre os resultados | 83 |
| 6.4.1 | Comparação entre os resultados do Exp1 e Exp2..... | 83 |
| 6.4.2 | Comparação entre os resultados Exp 2 e (Faria et al., 2009) | 85 |
| 6.5 | Resultados. Estratégia de Negociação | 86 |
| 7. | Conclusões e Trabalhos Futuros | 89 |
| 7.1 | Visão Geral do Trabalho | 89 |
| 7.2 | Conclusões gerais | 90 |
| 7.3 | Limitações e problemas do trabalho | 92 |
| 7.4 | Sugestões de Trabalhos Futuros | 92 |
| | REFERÊNCIAS BIBLIOGRÁFICAS | 94 |
| | APÊNDICE A | 104 |

LISTA DE FIGURAS

| | |
|--|----|
| Figura 2.1 Arquitetura Geral - Fase de Aprendizagem | 8 |
| Figura 2.2 Arquitetura Geral – Fase Operacional..... | 8 |
| Figura 3.1 Média diária de Negócios na BmfBovespa..... | 22 |
| Figura 3.2 <i>Candlestick</i> e suas representações..... | 24 |
| Figura 3.3 Gráfico de <i>candlestick</i> : Ibovespa (Período Agosto a Novembro de 2011)... | 24 |
| Figura 4.1 Arquitetura geral dos processos de Classificação de textos..... | 31 |
| Figura 4.2 Esquema de abordagem <i>Filter</i> | 36 |
| Figura 4.3 Esquema de abordagem <i>Wrapper</i> | 37 |
| Figura 4.4 Representação <i>Bag of Words</i> | 41 |
| Figura 4.5 Arquitetura básica de uma rede neural e seus principais componentes | 46 |
| Figura 5.1 Etapas da Metodologia do sistema proposto..... | 50 |
| Figura 5.2 Série diária do fechamento do Ibovespa (fevereiro de 2010 a junho de 2011) | 53 |
| Figura 5.3 Exemplo de Títulos de Notícias | 54 |
| Figura 5.4 Gráfico comparativo | 67 |
| Figura 6.1 Cotações do fechamento do Ibovespa e Variação Relativa do Ibovespa..... | 71 |
| Figura 6.2 Gráfico da média de notícias diárias no período de | 72 |
| fevereiro de 2010 a junho de 2011 | 72 |
| Figura 6.3 Acurácia dos classificadores para Base1..... | 73 |
| Figura 6.4 Acurácia dos classificadores para Base2..... | 74 |
| Figura 6.5 Resultados redes MLP para modelo de duas classes. Gráfico superior está relacionado à Base1, enquanto gráfico inferior está relacionado à Base2..... | 78 |
| Figura 6.6 Resultados redes RBF para modelo de duas classes. O gráfico superior está relacionado à Base1, enquanto gráfico inferior se refere à Base2..... | 79 |
| Figura 6.7 Lucro acumulado por ação | 86 |
| Figura 6.8 Lucro por dia de negociação | 87 |

LISTA DE TABELAS

| | |
|--|-----|
| Tabela 4.1 - Matriz Confusão | 48 |
| Tabela 5.1 - Estatísticas básicas dos títulos das notícias no período de 23/02/2010 a 30/06/2011 | 55 |
| Tabela 5.2 - Estatísticas básicas das notícias após estratégias adotadas | 57 |
| Tabela 5.3 - Distribuição para modelo de três classes para três limiares diferentes | 58 |
| Tabela 5.4 - Representação final dos documentos após a conversão tabela Atributo-Valor | 63 |
| Tabela 6.1 - Resultado modelo de previsão para Base2 | 75 |
| Tabela 6.2 - Medidas de Avaliação no Exp1 | 76 |
| Tabela 6.3 - Resultado modelo de previsão para Base1 | 82 |
| Tabela 6.4 - Medidas de Avaliação no Exp2..... | 82 |
| Tabela 6.5 – Comparação entre os resultados Exp1 e Exp2..... | 85 |
| Tabela A.1 - Comparação das principais características dos sistemas propostos na literatura..... | 105 |

1. Introdução

A área relacionada ao mercado financeiro sempre atraiu e atrai a atenção de diversos pesquisadores. No cenário atual de um mundo dinâmico e altamente globalizado o mercado financeiro em geral é um importante meio de financiamento empresarial, assim como também um importante meio de investimento individual. Além disto, a complexidade intrínseca destes sistemas faz da sua compreensão um desafio ou teste para todas as ferramentas e técnicas de análises de sistemas complexos. Tudo isto em conjunto, traduz a importância que é dedicada ao estudo dos mercados em várias universidades e comunidades científicas de diferentes países.

Atualmente a globalização tem correlacionado todos os mercados mundiais fazendo com que o comportamento dos mesmos seja complexo e de difícil previsão. Por outro lado, o avanço da computação e das comunicações nos dias atuais aperfeiçoou os mercados financeiros tornando-os mais eficientes. Para o investidor bastam agora poucos comandos para, em tempo real, efetuar operações financeiras em qualquer parte do mundo. Sendo assim o capital financeiro se desloca de um lugar para outro com uma grande velocidade, o que produz uma alta volatilidade nos mercados fazendo com que o volume negociado seja bem elevado e volátil. Dentre os mercados financeiros que mais se desenvolvem podemos citar o mercado de ações os quais são uma fonte de captação de recursos financeiros permanentes para as empresas. Com o crescimento deste mercado surge cada vez mais a preocupação em se entender o comportamento do mesmo.

Nos últimos anos os mercados acionários do mundo inteiro têm sido exaustivamente estudados por técnicas de Mineração de Dados (dados quantitativos ou informações numéricas geralmente dispostos em tabelas), onde dados históricos das ações, taxa de juros, indicadores econômicos, dentre outros, foram utilizados para tentar entender e prever o comportamento futuro dos mesmos em diferentes países. Muitos destes trabalhos foram revisados em (Krollner *et al.*, 2010). Porém, é amplamente conhecido que notícias desempenham um papel fundamental nos mercados financeiros. Estas notícias, nos formatos textuais, geralmente são dados não estruturados e difíceis de quantificar. Ao contrário da análise baseada em dados estruturados (mineração de dados) os dados de notícias textuais contêm não só o comportamento do mercado (por

exemplo, o preço da ação subiu), mas também a possível causa deste comportamento (por exemplo, o preço da ação subiu devido à diminuição da taxa SELIC- Sistema Especial de Liquidação e de Custódia).

O advento da internet fez com que o volume de notícias econômicas disponíveis na *web* tenha crescido vertiginosamente e o surgimento de novas tecnologias tem aumentado a capacidade de coleta e armazenamento destes dados de notícias. O trabalho do investidor financeiro está cada vez mais difícil uma vez que precisa processar um grande volume de dados textuais de maneira rápida e eficiente para sua tomada de decisão. Neste sentido as técnicas de Mineração de Textos tem sido uma abordagem muito utilizada para auxiliar os investidores nesta tarefa, uma vez que permite extrair, agregar e classificar grandes volumes de dados de notícias eficientemente. A utilização de técnicas de Mineração de Textos em estudos envolvendo os vários mercados financeiros mundiais vem crescendo muito nos últimos anos e em todos estes trabalhos tem-se a confirmação de que o conteúdo de notícias econômicas tem um forte impacto sobre o preço das ações (Gidófalvi, 2001). Nestes estudos diferentes estratégias têm sido abordadas. Robertson (2009) classifica as estratégias mais comumente utilizadas em três grupos:

- 1 **Estudos de eventos:** Busca por correlações entre notícias e comportamento anormal de negociação depois que certo tipo de notícia se torna disponível para o mercado (Kalev *et al.*, 2004). Nesta abordagem os pesquisadores estão apenas preocupados com o número de notícias de um determinado tipo, que ocorre em um período específico, de modo que o conteúdo da notícia é ignorado.
- 2 **Classificação de Textos:** Diferente da estratégia citada acima nesta abordagem o comportamento anormal de negociação do mercado é previsto utilizando o conteúdo das notícias (Robertson *et al.*, 2007). Dados reais dos preços dos ativos são utilizados para classificar as notícias.
- 3 **Análise de Sentimento:** Nesta abordagem palavras ou frases positivas (ex: “melhor que esperado”, “subiu”) e palavras ou frases negativas (“menor que esperado”, “caiu”) são procuradas na notícia a fim de inferir o sentimento do autor (Tetlock, 2007). A teoria é que quando os investidores estão sujeitos a um grande volume de notícias com sentimento negativo, eles passam a ter uma visão negativa do ativo e inversamente ocorre o mesmo, ou seja, se houver um excesso

de notícias positivas os investidores assumem uma posição mais otimista do ativo.

Nas abordagens 1 e 2 são utilizados algoritmos de aprendizagem de máquina para categorizar as notícias. A ideia básica destas estratégias parte do princípio de que os investidores de uma maneira geral antes de tomar suas decisões leem cuidadosamente notícias econômicas e financeiras recentes para ter uma visão atualizada do mercado. Usando seus conhecimentos de como o mercado se comportou no passado em diferentes situações, estes investidores vão corresponder implicitamente à situação atual como situações no passado, semelhantes à atual. Sendo assim os algoritmos de aprendizagem de máquina são treinados para reproduzir este conhecimento humano e prever situações em um período futuro. Para isto, técnicas de pré-processamento de textos (como eliminação de *stopwords*, algoritmos de *stemming* e outras) precisam ser utilizadas para transformar as notícias em dados estruturados que possam ser utilizados por estes algoritmos. É importante ressaltar que todas estas técnicas citadas serão detalhadas nas próximas seções deste trabalho.

Apesar deste crescente interesse nos últimos anos pela aplicação de técnicas de Mineração de Textos em notícias para entender o comportamento futuro dos mercados financeiros, este tipo de estudo ainda é uma área emergente de pesquisa. É uma abordagem pouco utilizada até o momento devido à dificuldade de se extrair informações relevantes a partir de dados não estruturados. O objetivo desta dissertação é explorar esta área emergente de pesquisa e estudar o comportamento do mercado de ações brasileiro. A motivação para estudar este mercado está no fato de que o mesmo é um mercado que mais movimenta recursos financeiros no Brasil. Podemos citar também que o mercado de ações do Brasil é considerado um dos mercados de maior liquidez entre os mercados emergentes e ainda não explorado na literatura. Tudo isto, somado a expansão da área de Mineração de Textos, serviu de motivação para o desenvolvimento deste trabalho. Os objetivos gerais do mesmo são descritos nos parágrafos seguintes.

1.1 Objetivos

O mercado financeiro é um sistema altamente complexo, dinâmico e não linear. A área relacionada à previsão financeira é caracterizada por dados ruidosos, não estacionários, não estruturados, com alto grau de incertezas e relacionamentos

“escondidos”. Muitos fatores interagem nos mercados financeiros incluindo eventos políticos, condições da economia em geral, expectativas dos investidores, entre outros. Portanto, prever o movimento dos preços nos mercados financeiros é uma tarefa muito difícil. O objetivo principal desta dissertação é através da abordagem de Mineração de Textos e de métodos de análise estatística prever a tendência do comportamento futuro do índice (Ibovespa) do mercado acionário do Brasil (BmfBovespa¹ – Bolsa de Valores, Mercadorias e Futuros), utilizando para isto os títulos (*headlines*) de notícias financeiras dos principais sites econômicos em português. O uso dos títulos como entrada para o modelo proposto está em linha com a abordagem proposta por (Peramunetilleke *et al.*, 2001), onde os autores afirmam que os títulos possuem um vocabulário restrito contendo somente as informações relevantes. Cabe também destacar que, até o momento da redação desta dissertação, não foi encontrado na literatura técnica científica, outro estudo similar que tenha abordado o tema desta dissertação.

Dentre os objetivos secundários que são abordados neste trabalho citamos:

- Criação de uma base de dados de notícias textuais em português relacionados ao mercado de ações brasileiro (esta base de dados textual é conhecida na literatura como *corpus*) a ser utilizado, uma vez que estes dados não existem ou não estão disponíveis para estudo.
- Verificar a influência das notícias publicadas antes da abertura do pregão da BmfBovespa na volatilidade diária do mercado de ações brasileiro.
- Estudo e análise de diferentes medidas estatísticas na fase de pré-processamento dos textos.
- Verificar a capacidade preditiva dos modelos que fazem uso de palavras chave obtidas automaticamente por métodos estatísticos e de modelos que fazem uso de palavras chave fornecidas por especialista do mercado de ações.
- Comparação dos resultados com os obtidos em (Faria *et al.*, 2009) onde o mesmo mercado foi estudado com dados estruturados e métodos de Redes Neurais Artificiais e Alisamento Exponencial Adaptativo.

¹ A BmfBovespa foi criada em 2008 com a integração entre a Bolsa de Mercadorias e Futuros (BM&F) e a Bolsa de Valores de São Paulo (BOVESPA).

1.2 Descrição e organização do trabalho

O presente trabalho está dividido em sete capítulos, seguindo uma linha em que primeiramente é feita uma revisão teórica do trabalho (capítulos 2, 3 e 4). E nos últimos capítulos são apresentados a metodologia proposta (capítulo 5), os resultados no capítulo 6 e por fim as conclusões no capítulo 7.

Mais detalhadamente, os capítulos citados podem ser definidos da seguinte forma. No primeiro capítulo é feita uma breve introdução à dissertação, sua motivação e seus objetivos.

O segundo capítulo fornece uma visão geral da literatura sobre a utilização de mineração de textos na previsão de diferentes mercados financeiros. Alguns conceitos básicos sobre mineração de textos e mercados financeiros são apresentados dentro do contexto dos trabalhos revisados. Cabe ressaltar que em caso de dúvidas sobre as técnicas de mineração de textos e ou conceitos envolvendo o mercado financeiro é recomendado a leitura dos capítulos 3 e 4 primeiro.

No terceiro capítulo são apresentados alguns conceitos do mercado financeiro brasileiro e as teorias utilizadas na predição dos mesmos (*random walk* e hipótese do mercado eficiente), além de duas filosofias distintas que surgiram a partir destas teorias (análise fundamentalista e análise técnica). Este capítulo introduz ao leitor alguns detalhes sobre os métodos de análise do mercado de ações que permitem uma melhor compreensão desta dissertação.

No quarto capítulo é apresentada a mineração de textos para análise de notícias financeiras. As principais tarefas de mineração de textos são apresentadas, mas somente a tarefa de classificação de textos é detalhada, uma vez que o principal objetivo desta dissertação é a classificação de notícias financeiras. É feito também uma breve descrição dos algoritmos de classificação utilizados neste trabalho, *Naive Bayes* (Bayesiano Simples), *Support Vector Machine* – SVM (Máquina de Vetor de Suporte) e *Artificial Neural Network* (Rede Neural Artificial - RNA).

No quinto capítulo a implementação da metodologia utilizada para alcançar os objetivos desta dissertação é apresentada. Dois experimentos distintos são utilizados na implementação da metodologia proposta. Todas as tarefas executadas nestes experimentos são detalhadas, assim como as métricas de avaliação dos resultados. Por fim uma estratégia de negociação baseada no sistema de previsão proposto é implementada.

O sexto capítulo apresenta todos os resultados obtidos com os diferentes parâmetros de configuração dos modelos criados e suas análises. Uma comparação entre os resultados desta dissertação e os resultados obtidos em (Faria *et al.*, 2009) também faz parte deste capítulo, assim como os resultados da estratégia de negociação adotada.

No último capítulo as conclusões são apresentadas e discutidas e por fim sugestões de possíveis trabalhos futuros são recomendadas.

2. Revisão da Literatura – Principais Trabalhos

Existe uma grande quantidade de artigos direcionados às técnicas de Mineração de Dados e previsão do mercado de ações, entretanto o número de artigos relacionados à Mineração de Textos e previsão do mercado de ações ainda é pequeno. Sendo assim, neste capítulo são apresentados os principais trabalhos de pesquisas existentes nesta área e as diferentes técnicas aplicadas pelos autores nos modelos desenvolvidos.

2.1 Arquitetura geral dos modelos de previsão do mercado e mineração de textos

A maioria dos trabalhos encontrados na literatura segue um padrão em comum apresentando duas fases distintas (uma fase de aprendizagem e uma fase operacional). Estas fases se referem à classificação automática de textos² (na qual com certeza, requer a etapa de pré-processamento de textos como passo preliminar). Durante a fase de aprendizagem, textos e séries temporais de cotações históricas do mercado (preço das ações, valores de índices, etc.) são coletados para treinar um classificador. Durante o treinamento os algoritmos de classificação tentam capturar as estruturas inerentes nas amostras dos documentos pré-classificados. Os resultados desta fase são então utilizados para classificar novos documentos na fase operacional, ou seja, nesta fase os artigos publicados recentemente alimentam o classificador desenvolvido previamente para prever as tendências futuras ou outras características do mercado. Sendo assim a arquitetura apresentada nas figuras 2.1 e 2.2 refere-se à fase de aprendizagem e fase operacional respectivamente de um modelo de previsão da tendência do mercado de ações. Esta arquitetura geralmente é a base para os modelos encontrados na literatura.

² Detalhes sobre a etapa de classificação automática de textos são apresentados no capítulo 4 desta dissertação.

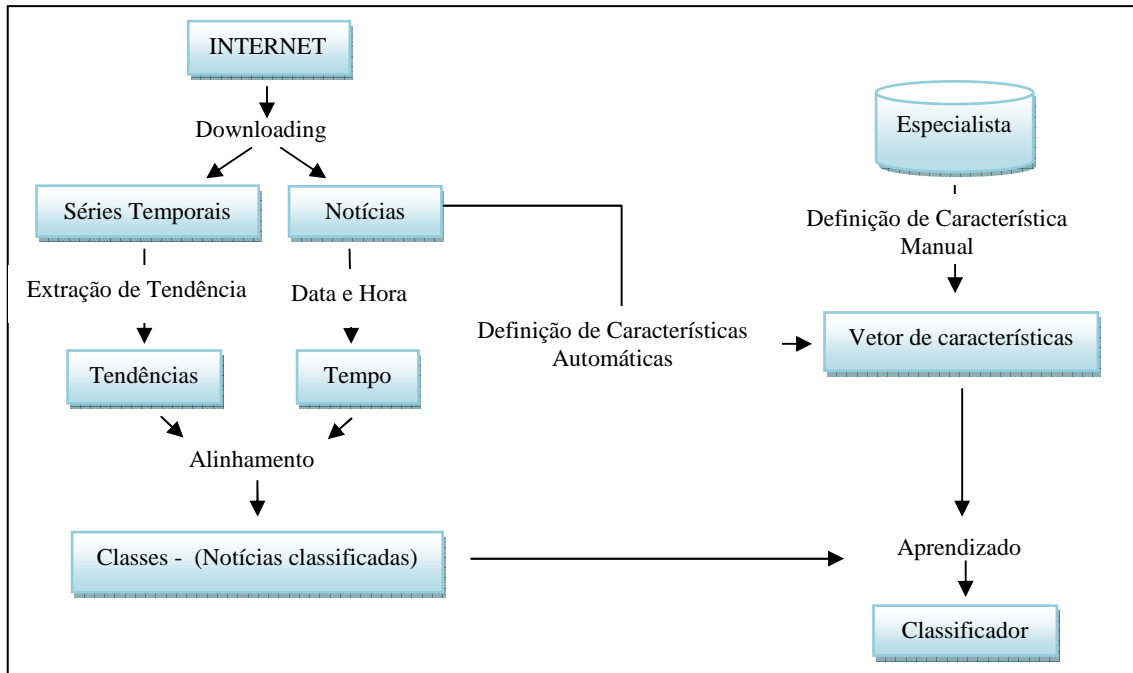


Figura 2.1 Arquitetura Geral - Fase de Aprendizagem

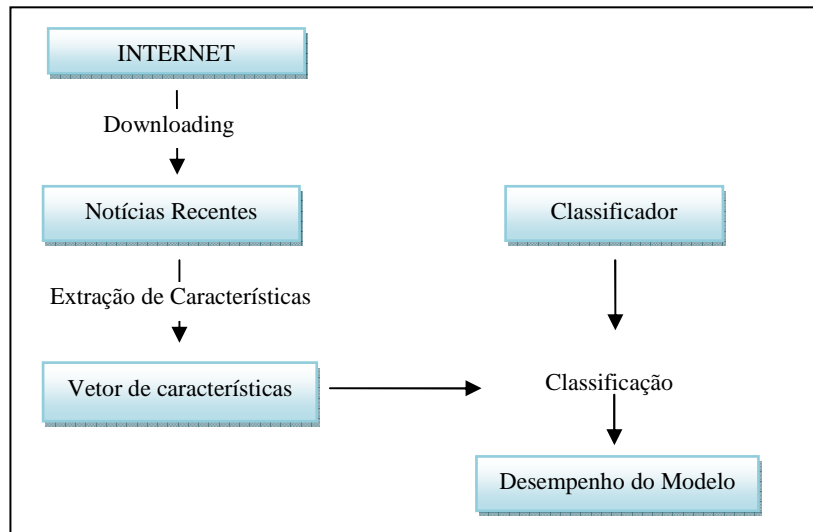


Figura 2.2 Arquitetura Geral – Fase Operacional

Na figura 2.1 acima textos e séries temporais são recuperados a partir da internet e pré-processados de forma tal a serem utilizados pelos algoritmos de classificação. O primeiro passo é classificar as notícias de acordo com o efeito que a mesma exerce sobre o mercado de ações (por exemplo, sobe, cai ou se mantém estável). Para isto são utilizadas as tendências extraídas a partir das séries temporais e data e hora das notícias.

O próximo passo está relacionado à definição das características³ que pode ser feita automaticamente por método de inferência ou podem ser fornecidas por um especialista. Uma vez definida, as mesmas são transformadas em um vetor de características e este vetor é então utilizado pelo algoritmo de classificação. Este algoritmo conforme citado anteriormente, através de um processo de aprendizado (treinamento) é capaz de capturar as estruturas inerentes nas amostras dos documentos pré-classificados. Já na fase operacional (figura 2.2) o classificador induzido na fase de aprendizagem é utilizado com novas notícias (também pré-processadas e transformadas em vetor de características) e uma medida de desempenho é utilizada para avaliar o modelo de previsão.

Conforme citado anteriormente o modelo citado nas figuras se refere à previsão da tendência do mercado. Esta escolha foi feita primeiramente porque está em linha com o principal objetivo deste trabalho (que é a previsão da tendência do mercado de ações da BmfBovespa) e também porque muitos dos trabalhos encontrados na literatura afirmam que do ponto de vista dos investidores de Bolsa de Valores as tendências das séries temporais financeiras são muito mais informativas do que o preço exato das ações ((Lavrenko *et al.*, 2000); (Gidófalvi, 2001)). Entretanto, conforme poderá ser observado na próxima seção, alguns autores não concordam com esta abordagem e preferem trabalhar diretamente com o preço das ações (Schumaker *et al.*, 2009). É importante destacar que para estes estudos em específico o módulo de extração de tendências apresentado na fase de aprendizagem da figura 2.1 não se faz necessário no modelo.

2.2 Resumos dos principais trabalhos

O uso de mineração de textos para prever o mercado financeiro teve início no final dos anos 90. Um dos primeiros trabalhos envolvendo este tema foi publicado em (Wuthrich *et al.*, 1998). Neste artigo os autores desenvolveram um sistema online cujo objetivo foi prever os índices diários das principais Bolsas de Valores mundiais (Ásia, Estados Unidos e Europa) utilizando para isto notícias publicadas nos jornais “*Financial Times*”, “*Reuters*”, e “*Wall Street Journal*”. Os autores utilizaram mais de 400 sequências de palavras chave que foram fornecidas por especialistas do mercado e

³ Não será feita distinção entre características ou palavras chave nesta dissertação, pois ambas representam o mesmo conceito dentro da área de mineração de textos.

que poderiam ter um impacto sobre o mercado de ações. Diversas técnicas de aprendizagem de máquina como redes neurais artificiais, *K-Nearest Neighbor* (K-NN) e algoritmos baseado em regras foram utilizados para produzir a previsão dos índices em estudo. Estas previsões estavam disponíveis diariamente e em tempo real no site www.cs.ust.hk/~beat/Predict⁴ às 07h45min da manhã (horário de *Hong Kong*). Conseqüentemente os mercados de *Tóquio*, *Hong Kong*, *Singapura* e principais mercados Asiáticos tinham acesso às previsões antes de começarem suas negociações em suas respectivas Bolsas de Valores. Por fim os autores concluíram que a utilização de informação textual juntamente com séries temporais financeiras (valores diários dos índices dos mercados) aumenta a qualidade das variáveis de entrada dos modelos.

A partir deste trabalho pioneiro, diversos outros trabalhos foram publicados nos últimos anos. Um resumo contendo as principais diferenças (objetivos, técnicas adotadas, mercados envolvidos, desempenho, etc.) dos primeiros artigos publicados sobre previsão do mercado de ações e previsão da taxa de câmbio pode ser visualizado na tabela apresentada no Apêndice A, que está organizada em quatro seções: objetivo do sistema (onde é fornecida uma ideia do sistema desenvolvido), parâmetros de mineração de textos (onde são apresentados os parâmetros da mineração de textos utilizados nos sistemas desenvolvidos), dados de entrada (são apresentados os dados utilizados) e por fim na última seção, teste (onde é apresentado um resumo dos principais desempenhos reportados pelos autores). Uma descrição mais detalhada sobre os sistemas citados nesta tabela e suas propriedades é apresentada a seguir, assim como outros sistemas citados na literatura também, porém mais recentes.

Lavrenko *et al.* (2000) e Lavrenko *et al.* (2000a) são trabalhos publicados pelos mesmos autores com diferentes títulos, mas com conteúdo muito semelhante. Ambos tiveram a finalidade de prever a tendência dos preços de diferentes ações dos Estados Unidos, i.e. o movimento *intraday*, analisando notícias publicadas no site da *Yahoo* (seção de finanças). Um sistema chamado *Enalist* foi desenvolvido pelos autores para esta finalidade. O *Enalist* opera correlacionando o conteúdo das notícias com tendências em séries temporais financeiras. Os autores primeiramente segmentaram as séries temporais dos preços das ações em pequenas janelas de tendências utilizando um algoritmo de regressão linear por etapas (*piecewise linear regression*). Detalhes deste algoritmo podem ser encontrados em (Keogh *et al.*, 2001). Em seguida as tendências

⁴ Não está mais disponível.

foram discretizadas onde “*labels*” (ou rótulos) foram atribuídas aos segmentos baseados em suas características (tamanho, inclinação, intercepto e coeficiente r^2) utilizando para isto um algoritmo de agrupamento aglomerativo (Everetti, 1993). O próximo passo do trabalho foi alinhar cada tendência com as notícias. Uma notícia foi associada a uma tendência se a hora da publicação da notícia foi h horas antes do começo de uma janela de tendência. Os autores afirmam que o parâmetro h que forneceu melhores resultados variou entre 5 e 10 horas. Diferentemente do sistema desenvolvido por (Wuthrich *et al.*, 1998), as palavras chave neste trabalho foram determinadas automaticamente através da medida TF-IDF (*Term Frequency - Inverse Document Frequency*). Um classificador *Naive Bayes* foi treinado para identificar padrões de comportamento entre as notícias e suas respectivas tendências. Por fim os autores fizeram uma simulação do mercado para avaliar o modelo proposto e concluíram que o *Enalist* quando comparado a um modelo randômico obteve melhores resultados (ver tabela Apêndice A). É importante salientar que neste trabalho os autores afirmam que deve existir um período, (chamado pelos autores de h , e definido como 5 horas) para que o mercado absorva qualquer tipo de informação publicada. Muitos pesquisadores admitem que o mercado demora um tempo para digerir tais informações, entretanto um longo período como este citado pelos autores contradiz muitas teorias econômicas, como, por exemplo, a hipótese do mercado eficiente.

Thomas *et al.* (2000) também publicou um trabalho no qual duas abordagens foram utilizadas para prever o mercado financeiro. A primeira abordagem utilizou um classificador de texto utilizando a técnica de Máxima Entropia (Nigan *et al.*, 1999) para determinar o impacto dos comunicados da *web* sob os preços das ações. Na segunda abordagem regras de negociação simples foram construídas utilizando algoritmos genéticos e baseados no volume negociados das ações em estudo, no número de comunicados e também no número total de palavras postado por dia na *web*. Os autores afirmaram estarem mais interessados no lucro gerado pelas regras de negociação do que nas previsões propriamente. Sendo assim, os mesmos reportaram que as duas abordagens produziram lucros, mas que a integração das duas abordagens aumentou em 30% a lucratividade. Entretanto nenhuma análise ou avaliação dos modelos propostos foi apresentada neste trabalho.

Outro trabalho visando à previsão da tendência dos preços das ações por meio de notícias financeiras foi publicado por (Gidófalvi, 2001). Neste artigo os autores

inicialmente definiram um intervalo de tempo que foi chamado de “*janela de influência*”. Esta janela de influência é o período de tempo no qual a notícia tem um efeito nos preços das ações, sendo que a mesma é caracterizada por um limite inferior e um limite superior a partir do momento t da publicação de uma notícia. Neste sentido os autores afirmaram que após muitos experimentos um intervalo de 20 minutos foi o escolhido. Isto significa então que uma notícia publicada em um período t influenciará o preço das ações 20 minutos antes de sua publicação (t) e 20 minutos após a sua publicação (t). Em seguida as notícias foram classificadas em três categorias, sendo que a primeira categoria consistiu em notícias que aumentaram os preços das ações em pelo menos 0,2% durante a janela de influência, a segunda categoria, em notícias que diminuiram os preços das ações em pelo menos 0,2% e as notícias restantes foram classificadas na terceira categoria. Assim como em (Lavrenko *et al.*, 2000) as palavras chave do modelo proposto pelos autores foram definidas automaticamente e um classificador *Naive Bayes* foi treinado para prever a qual categoria uma dada notícia pertence. Os autores concluíram que os resultados encontrados não foram satisfatórios (conforme tabela Apêndice A), porém eles afirmaram que encontraram uma forte correlação entre as notícias e o comportamento do preço das ações durante os 20 minutos que antecedem a publicação da notícia e os 20 minutos seguintes. Este resultado também discorda da hipótese do mercado eficiente e sugere que é possível obter lucros em um curto intervalo de tempo.

Peramunetilleke *et al.* (2002) desenvolveu um sistema em colaboração com o banco UBS (União dos Bancos Suíços) cujo objetivo foi prever o movimento *intraday* das taxas de câmbio entre o Dólar (EUA) e o Yen (Japão) utilizando *headlines* (títulos) ao invés de utilizar todo o conteúdo da notícia. Assim como em (Wuthrich *et al.*, 1998) os autores utilizaram um dicionário de palavras chave fornecidas por especialistas (*traders* da UBS), onde mais de 400 palavras chave contendo de duas a cinco palavras combinadas ou não com o operador *and* foram fornecidas e também utilizaram as mesmas técnicas para classificar as notícias. Por fim os autores concluíram que os resultados encontrados foram superiores quando comparados a um *trader randômico* e iguais quando comparados a resultados fornecidos pelos *traders* profissionais da UBS. Eles reportaram também que a contribuição deste trabalho é a modelagem inteligente de um problema de previsão permitindo o uso de conteúdo textual.

Fung *et al.* (2002) introduziu em seu trabalho outra metodologia para prever tendências das ações utilizando notícias. O que o distingue dos trabalhos anteriores é o fato dos autores investigarem o impacto imediato das notícias sobre as séries temporais baseado na hipótese do mercado eficiente. Diferentemente do modelo proposto por (Lavrenko *et al.*, 2000) onde um intervalo fixo de 5 horas foi utilizado para alinhar as tendências e notícias, neste trabalho nenhum período fixo é necessário, pois a hipótese do mercado eficiente afirma que o mercado atual reflete a assimilação imediata de toda a informação disponível. Logo um intervalo de tempo longo é normalmente impossível. Os dados e notícias utilizados neste trabalho se referem a 614 ações listadas na Bolsa de Valores de *Hong Kong* durante sete meses consecutivos. Um algoritmo de regressão linear por etapas (*piecewise linear regression*) foi utilizado para encontrar as tendências nas séries temporais. As tendências segmentadas foram então agrupadas em duas categorias (sobe, desce) de acordo com a inclinação das tendências e do coeficiente r^2 . Em seguida as notícias foram alinhadas com as tendências utilizando uma extensão do algoritmo *k-Means* proposto por (Kaufman *et al.*, 1990). Um esquema diferenciado para atribuir pesos às palavras mais relevantes das notícias (palavras chave) foi proposto pelos autores. Neste esquema, as palavras chaves que ocorrem em um determinado agrupamento e não ocorre ou raramente ocorre no outro agrupamento tem maior importância, sendo assim recebe uma pontuação maior. O algoritmo SVM foi utilizado para aprender os padrões encontrados entre as notícias e suas respectivas tendências. Para avaliar o sistema proposto, os autores fizeram uma simulação do mercado utilizando uma estratégia de negociação simples baseada no teste *Buy-and-Hold*⁵. Apesar de não colocarem os valores dos lucros obtidos com o modelo proposto conforme os trabalhos citados anteriormente, os autores afirmaram que os lucros obtidos na abordagem proposta foram rentáveis.

Em 2003 Fung *et al.* propôs outro modelo muito semelhante ao trabalho anterior, porém ao invés de utilizar séries temporais simples, os autores utilizaram séries temporais múltiplas. Eles ressaltaram que muitos dos trabalhos existentes na literatura estão preocupados somente com séries temporais simples e que o relacionamento e influência entre as diferentes ações são ignoradas. A ideia básica deste trabalho foi

⁵ *Buy and Hold* é uma estratégia de investimento a longo prazo, baseado no fato de que, a longo prazo os mercados financeiros apresentam uma boa taxa de retorno mesmo em períodos de alta volatilidade ou durante colapsos financeiros.

explorar as séries temporais múltiplas e prever seus movimentos utilizando somente informação textual (assim como o trabalho anterior dos próprios autores) baseado na hipótese do mercado eficiente. A estrutura deste modelo tem um passo adicional quando comparado ao trabalho anterior que é a descoberta do relacionamento entre as séries temporais e o alinhamento das notícias em múltiplas séries temporais. Detalhes deste trabalho não estão disponíveis aqui, uma vez que o foco desta dissertação está apenas no estudo de séries temporais simples. Todavia, cabe salientar que os lucros obtidos por este modelo foram quase que o dobro quando comparado aos lucros obtidos nos modelos anteriores de outros autores.

Um sistema chamado *NewsCATS* (*News Categorization and Trading System* ou Sistema de Negociação e Classificação de Notícias) foi desenvolvido por (Mittermayer, 2004), cujo o objetivo foi prever a tendência dos preços das ações no instante imediato em que a notícia foi publicada. O sistema desenvolvido pelos autores consistiu de três componentes. Sendo que o primeiro componente (mecanismo de pré-processamento de documentos) recupera informações relevantes nas notícias disponíveis na *web* através da aplicação de técnicas de pré-processamento de textos. Já o segundo componente (mecanismo de classificação) classifica as notícias em categorias pré-definidas. E por fim estratégias de negociação são determinadas pelo terceiro componente (mecanismo de negociação) através da classificação anterior. O *NewsCATS* utiliza um arquivo de notícias e um arquivo dos preços *intraday* de todas as ações listadas na *S&P500* (Bolsa de Valores dos Estados Unidos) no período de 01-01-2002 a 31-12-2002 a cada 60 minutos. Com estes arquivos o *NewsCATS* é capaz de aprender um conjunto de regras que permite classificar as notícias automaticamente em um número definido de categorias (ou classes) através do mecanismo de classificação. Sendo que cada uma destas categorias está associada a um impacto específico no preço das ações (por exemplo, a ação subiu ou caiu). Dependendo do resultado fornecido pelo mecanismo de classificação, o mecanismo de negociação produz sinais de negociação que podem ser executados via um *broker*⁶ online ou por outros intermediários. O mecanismo de pré-processamento de documentos inclui extração de palavras chaves (*stem* e eliminação de *stop words*) e seleção de características através da *TF*(*Term Frequency*), *IDF*(*Inverse Document Frequency*) ou *TF-IDF*(*Term Frequency - Inverse Document Frequency*) como medida estatística. A representação do documento pode ser realizada com uma

⁶ Broker é um sistema de negociação de ações online.

medida booleana da frequência ou com a *TF*, *IDF* ou *TF-IDF*. Vale ressaltar que todas estas técnicas de pré-processamento de textos são detalhadas no capítulo 4 deste trabalho. A saída deste pré-processamento é encaminhada ao mecanismo de classificação onde as notícias são classificadas em três classes (*good news*, *bad news* e *no movers*). A classificação automática das notícias foi feita utilizando o algoritmo de classificação SVM. Finalmente o mecanismo de negociação gera os sinais de negociação das ações que são as saídas do sistema *NewsCATS*. Os autores por fim reportaram que as estratégias de negociação fornecidas pelo modelo proposto superaram significativamente um *trader randômico* comprando e vendendo ações randomicamente imediatamente depois da publicação das notícias.

Yu *et al.* (2005) estudaram o mercado de *commodities*, mais precisamente o preço do petróleo. Técnicas de mineração de textos juntamente com a teoria *rough set* foram utilizadas para prever a tendência do mercado de petróleo. O sistema desenvolvido pelos autores foi dividido em dois módulos. No primeiro módulo, fatores (econômicos, políticos, militares, desastres naturais, especulação, entre outros) que poderiam impactar o preço do petróleo foram buscados na internet a fim de se construir um repositório de dados. Estes dados foram então processados utilizando as técnicas de mineração de textos para gerar conhecimento. Este processo de mineração de textos foi dividido em quatro fases: coleta de documentos relevantes, pré-processamentos dos documentos, extração de palavras chaves e por fim mineração de dados e geração de conhecimento. No segundo módulo a teoria *rough set* proposta por (Pawlak, 1982) foi utilizada para reduzir inconsistência nos padrões encontrados no módulo anterior. Os objetivos principais aqui foram checar dependências (parciais ou totais) entre atributos, reduzir os atributos, analisar a significância dos atributos e gerar regras de decisão. Combinando estes dois módulos regras e padrões (conhecimento) foram gerados para prever a tendência do mercado de petróleo. Para avaliar a habilidade de previsão do modelo proposto, os autores fizeram uma comparação de seus resultados com métodos convencionais (modelos estatísticos e modelos de séries temporais) e redes neurais artificiais, onde ficou comprovado que o modelo proposto apresentou resultados superiores. Sendo assim os autores concluíram que a abordagem proposta por eles é uma alternativa promissora aos métodos convencionais de previsão de tendência do petróleo.

Em 2006 o *NewsCATS* proposto por Mittermayer (2004) foi reformulado e um novo trabalho foi publicado (Mittermayer *et al.*, 2006a). Neste trabalho os autores também utilizaram os mesmos componentes citados no trabalho anterior (mecanismo de pré-processamento de documentos, mecanismo de classificação e mecanismo de negociação), porém com algumas modificações. A fase de seleção de características além do procedimento automatizado, também contava com um dicionário criado pelos autores contendo palavras simples, frases e sequências de palavras chaves que poderiam influenciar o preço das ações. A fim de se reduzir o ruído presente nos dados, o arquivo de notícias foi modificado com a implementação de uma heurística para separar as notícias com tópicos que realmente poderiam afetar o preço das ações. E o arquivo de séries temporais passou a conter os preços *intraday* das ações a cada 15 segundos, o que aumentou a frequência dos dados permitindo assim uma avaliação de desempenho mais realista. Diferentes parâmetros (tamanho do conjunto de características, representação do documento, algoritmos de classificação, entre outros) foram testados e analisados tornando o *NewsCATS* mais robusto. Por fim os autores reportaram que os lucros obtidos foram de 0,29%, o que segundo os mesmos é uma melhoria notável quando comparado a outros sistemas existentes na literatura (para comparação de alguns destes trabalhos ver tabela apêndice A).

Schumaker *et al.* (2006) examinou as notícias financeiras sob o ponto de vista de três diferentes representações textuais. Foram utilizadas as representações *Bag of words*, *Noun Phrases* e *Named Entities* para prever o preço das ações vinte minutos após a publicação de uma notícia. Diferentemente de muitos artigos citados anteriormente, os autores deste trabalho focalizaram seus objetivos em prever o preço das ações e não a tendência dos preços e também na utilização de outras representações textuais, dado que a abordagem “*bag of words*” é a abordagem padrão mais utilizada na literatura. Para atingir seus objetivos os autores desenvolveram um sistema chamado *AZFinText* (*Arizona Financial Text System* ou Sistema de Texto Financeiro do Arizona). Neste sistema cada notícia foi representada utilizando as três técnicas de representação textual citada anteriormente e um conjunto de palavras chaves por fim foram armazenadas em um banco de dados. As cotações das ações também são armazenadas de minuto a minuto. Quando uma notícia é publicada o sistema estima qual deve ser o valor da ação após 20 minutos. Para isto uma regressão linear utilizando o preço das ações foi calculada 60 minutos antes da notícia ser publicada e o preço das ações 20 minutos no

futuro é estimado. O período de 20 minutos foi escolhido seguindo o trabalho de (Gidóvalfi, 2001). Por fim notícias e preço das ações foram processados pelo algoritmo *SVR (Support Vector Regression)* que é derivado do algoritmo SVM, porém sem o aspecto de classificação, ou seja, o *SVR* trabalha com análise de valores discretos. Foi utilizado um kernel linear e validação cruzada em 10 ciclos. Três métricas foram utilizadas para avaliar o modelo proposto e em todas as avaliações foi comprovado que os resultados obtidos com as três representações textuais foram superiores quando comparados ao método de regressão linear. Os autores também demonstraram que das três representações textuais, *Noun Phrases* foi a que obteve melhores resultados.

Outra abordagem que também vem sendo utilizado na literatura para prever o mercado de ações é a combinação de indicadores técnicos e notícias financeiras. Em (Zhai *et al.*, 2007) sete indicadores técnicos calculados para cada dia de negociação a partir dos preços históricos dos últimos cinco dias das ações, juntamente com notícias relacionadas diretamente com a ação e notícias do mercado em geral (relacionadas indiretamente com a ação) foram utilizados para prever a tendência diária dos preços das ações da *BHP Billiton Ltda.*⁷, da Bolsa de Valores da Austrália no período de março de 2005 a maio de 2006. Os autores desenvolveram cinco modelos utilizando o algoritmo SVM e cinco variáveis de entrada diferentes. Sendo que no primeiro modelo foram utilizados como variáveis, apenas indicadores técnicos, no segundo modelo, notícias diretamente envolvidas com as ações, no terceiro modelo, as notícias do mercado em geral, no quarto modelo foram utilizados a combinação de notícias (diretas e indiretas) e por fim a combinação de indicadores técnicos e notícias no último modelo. Os autores reportaram que a acurácia do modelo proposto foi comparável à obtida por (Kim, 2003) onde somente indicadores técnicos foram utilizados (acurácia de 58,8%), porém o desempenho do último modelo (combinação de indicadores técnicos e notícias) foi muito superior (acurácia de 70,1%). Os autores também reportaram uma simulação do mercado para avaliar a lucratividade do sistema. Para este fim três conjuntos de estratégias foram utilizados para os diferentes modelos propostos e os resultados demonstraram que os lucros obtidos pelo último modelo foram superiores quando comparados aos outros modelos e também quando comparados a um sistema de

⁷ *BHP Billiton* se refere a maior empresa de mineração do mundo, criada a partir da fusão da empresa australiana *Broken Hill Proprietary Company* (BHP) com a empresa inglesa, *Billiton*.

negociação que emprega uma estratégia randômica, ou seja, compra e vendas de ações realizadas randomicamente no mesmo período estudado.

Em 2009 o sistema *AZFinText* proposto por (Schumaker *et al.*, 2006) foi modificado e um novo trabalho visando outros objetivos foi publicado (Schumaker *et al.*, 2009). Os objetivos dos autores neste trabalho foi verificar qual o efeito que o particionamento GICS (*Global Industry Classification Standard*)⁸ de notícias poderia ter sobre a previsão do preço das ações. Os autores reportaram que neste sentido esta pesquisa é totalmente nova, dado que a maioria dos trabalhos na literatura utiliza em seus modelos notícias financeiras universais ou específicas para o objetivo-alvo. O segundo objetivo foi determinar qual a capacidade preditiva do modelo quando comparado à previsão feita por especialistas do mercado e gestores de fundos quantitativos.

Para testar o efeito do particionamento GICS vários modelos foram desenvolvidos para cada classificação GICS (Setor, Grupo de Indústria, Indústria e Subindústria) e também para cada ação específica. Os mesmos parâmetros que apresentaram os melhores resultados no modelo anterior dos autores foram utilizados, ou seja, representação textual *Proper Nouns*, algoritmo *SVR*, previsão 20 minutos à frente, assim como as mesmas métricas de avaliação. Os resultados demonstraram que as ações do particionamento de Setor foram as que tiveram melhor desempenho. Quanto à comparação entre o modelo proposto e especialistas do mercado, os autores concluíram que os resultados obtidos pelo sistema desenvolvido foram superiores. Já em relação à comparação com os fundos quantitativos, os resultados obtidos pelos autores superaram os resultados obtidos por 6 dos 10 melhores fundos de 2005 e sendo que para fundos que monitoram os mesmos títulos o sistema proposto teve um retorno de 2% a mais que o melhor desempenho dos fundos.

Muitos dos trabalhos citados acima e também disponíveis na literatura estão preocupados com a previsão da tendência de mercado a curto prazo (de hora em hora, a cada 15 minutos, diário, etc.). Kiyoshi *et al.* (2010) diferentemente destes trabalhos analisou o mercado de títulos do Japão através das técnicas de mineração de textos, porém visando a previsão da tendência do mercado a longo prazo (mensalmente). Segundo os autores dois pontos são importantes quando se está investigando a previsão

⁸ Sistema GICS (*Global Industry Classification Standard*) do banco *Morgan Stanley*, que possui quatro níveis hierárquicos: setor, grupo de indústria, indústria e subindústria.

de tendência do mercado a longo prazo e técnicas de mineração de textos. O primeiro é a utilização de dados textuais com conteúdo e formatos apropriados. Os autores afirmam que neste sentido foram utilizados relatórios mensais sobre desenvolvimentos econômicos e financeiros recentes do *BOJ (Bank of Japan)*. Este relatório analisa a situação financeira e econômica japonesa de um ponto de vista macro. Três razões que tornaram adequadas a utilização deste relatório para uma análise de mercado a longo prazo foram citadas pelos autores: 1) É regularmente lançado podendo portanto ser rastreado quaisquer mudanças temporais nos dados textuais. 2) Quase todos os *traders* institucionais do mercado japonês prestam atenção a estes relatórios, tendo seus comportamentos de negociação muitas das vezes afetados pelos mesmos, porque eles refletem a decisão inicial tomada pelo *BOJ*. 3) Tem um formato regular, sendo assim um documento é facilmente comparável a outro em diferentes pontos ao longo do tempo.

Já o segundo ponto está relacionado ao método para associar dados textuais e séries temporais. Para esta finalidade foi proposto pelos autores o método *CPR* que consiste de três passos: 1 - Análise de ocorrência de palavras chaves, 2 - Análise de componentes principais das mudanças temporais nas palavras chaves e 3 - Análise de regressão dos dados dos preços utilizando componentes principais.

Para testar o modelo proposto duas regras de simulação foram propostas, uma para comparar os preços dos títulos e outra para comparar a tendência dos mesmos. Por fim os autores reportaram que os resultados obtidos pelo método proposto foram superiores quando comparados a outros métodos de previsão disponíveis na literatura. E com isto concluíram que dados textuais podem ser utilizados para análise de mercado, não só a curto prazo, mas também a longo prazo.

De todos estes trabalhos, é possível concluir que este tema está em aberto. O número de modelos propostos para previsão do mercado financeiro utilizando dados não estruturados (dados textuais) não é comparável ao número de modelos propostos onde dados estruturados são utilizados para esta mesma finalidade. Entretanto alguns pesquisadores têm sido bem sucedidos até certo ponto, na previsão do mercado financeiro utilizando dados não estruturados. A maioria destes pesquisadores tem comprovado que as regras de negociação geradas por seus modelos são mais rentáveis do que um *trader randômico*. Várias técnicas e abordagens assim como diferentes atributos, nos quais é possível citar tipos de dados, técnicas de pré-processamento,

algoritmos de classificação, entre outros, também estão sendo utilizadas para garantir uma maior acurácia nos modelos desenvolvidos.

Por tudo o que foi dito anteriormente, além de outras razões, fazem com que este tema de pesquisa mostre um grande crescimento na comunidade científica mundial recebendo um número cada vez maior de publicações e congressos.

3. Mercado de Ações no Brasil

Mauro, (2001) define Bolsa de Valores como um clube (ou grupo sem fins lucrativos) de corretores de valores, intermediários e pessoas interessadas ou ligadas à compra e venda de ativos financeiros. Os negócios e transações realizadas por estas pessoas e grupos são públicas, transparentes e podem ser acompanhadas pelos meios de difusão em massa. O mercado de ações do Brasil está situado na cidade de São Paulo e é conhecido atualmente pela sigla BmfBovespa formada em 2008, a partir da integração das operações da Bolsa de Valores de São Paulo e da Bolsa de Mercadorias e Futuros. Como principal instituição brasileira de intermediação para operações do mercado de capitais, a companhia desenvolve, implanta e provê sistemas para a negociação de ações, derivativos de ações, títulos de renda fixa, títulos públicos federais, derivativos financeiros, moedas à vista e commodities agropecuárias (BMF, 2012).

O mercado brasileiro ou BmfBovespa é um dos mercados de maior liquidez dentro dos mercados emergentes. Aproximadamente sete bilhões de reais são movimentados todos os dias na BmfBovespa e a média de negócios vem aumentando consideravelmente a cada ano, conforme pode ser visualizado na figura 3.1. O crescimento físico de investidores domésticos também é outro fator responsável pela evolução dos negócios da bolsa. Nos dias atuais o número de investidores que tem CPF cadastrado como investidor chega perto de 600 mil pessoas. Estima-se que até 2014 este número chegue a 5 milhões de investidores (BMF, 2012).

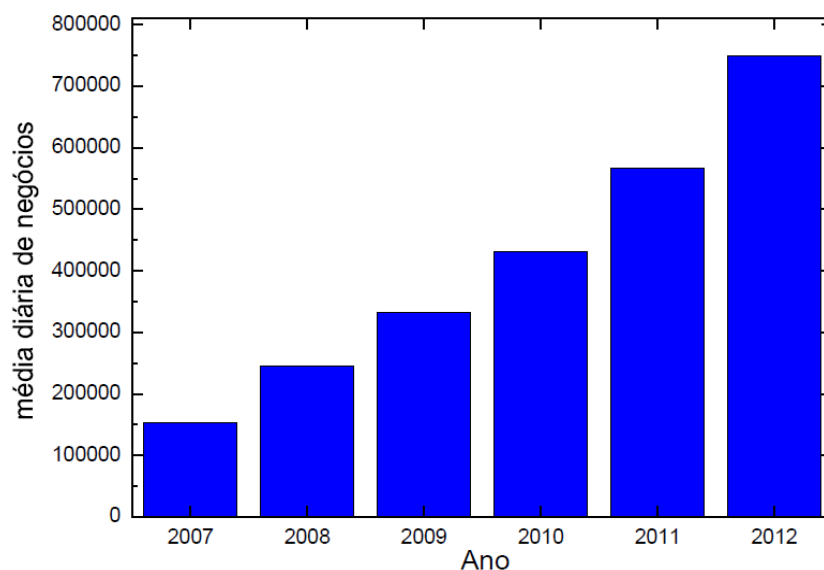


Figura 3.1 Média diária de Negócios na BmfBovespa

(2012 compreende aos meses de janeiro e fevereiro 2012) - Fonte: (BMF, 2012).

Na Bolsa de Valores são negociadas as ações de diversas empresas que possuem capital aberto. Ação é uma unidade de títulos emitidos que representa a menor fração do Capital Social da empresa. O investidor ao adquirir uma ação na Bolsa de Valores ele se torna um sócio da companhia, mesmo que em proporção pequena em relação ao controlador (sócio majoritário) e os poderes a ele atribuídos são limitados pelo tipo de ação que comprou e também pela quantidade de ações que possui.

Existem dois tipos de ações (ordinárias e preferenciais), que designam direitos e poderes diferentes a seus titulares. Por exemplo, o acionista tem preferência no recebimento de dividendos (percentual sobre o lucro da companhia), porém não tem direito ao voto nas decisões do Conselho Administrativo no caso das ações preferenciais (PN). Já as ações ordinárias (ON) concedem a aqueles que são acionistas o direito a voto nas decisões da empresa, mas não têm preferência na distribuição de resultados.

3.1 Características do Mercado de Ações

O preço de uma ação é determinado pela compra e venda das mesmas. Desde que haja procura no mercado, as ações podem ser convertidas em dinheiro a qualquer momento por intermédio de uma Sociedade Corretora e através da negociação na Bolsa de Valores. O valor do preço das ações das empresas que tem suas ações negociadas na

BmfBovespa estão sujeitas à lei da oferta e da procura. As ações de uma empresa podem ser muito negociadas (chamadas de *blue chips*), ou pouco negociadas (chamadas de *small caps*). Para avaliar o sentimento do mercado como um todo é necessário um indicador que reflita o comportamento geral do mercado. No mercado de ações brasileiro este indicador (ou índice) é conhecido como Ibovespa e representa o desempenho médio dos preços das ações mais negociadas na BmfBovespa ponderado por um fator proporcional ao volume negociado das mesmas. Vale ressaltar que este é o indicador a ser estudado nesta dissertação.

O Ibovespa conjuntamente com o preço das ações das empresas negociadas na BmfBovespa oscila durante um dia de negociação. Estas oscilações dependem de muitos fatores, como notícias, eventos (políticos, econômicos, *etc.*) e interesse dos investidores. Um dia de grande procura por ações leva ao aumento dos preços das mesmas e conseqüentemente ao aumento do Ibovespa.

A sequência diária de preços de qualquer ação pode ser representada através de uma série temporal. Na série cada dia pode ser representado pelo desenho de uma estrutura chamada de *candlestick*, que representa graficamente a variação de preços de uma determinada ação em um determinado período. Nele está representado o preço de abertura (ou valor no caso do Ibovespa) que compreende ao preço inicial ou ao primeiro negócio fechado com as ações de uma determinada empresa na abertura do pregão⁹, preço de fechamento que compreende ao último negócio (ou *call* de fechamento) efetuado no horário regular do pregão e por fim os valores máximos e mínimos que compreendem aos preços máximos e mínimos respectivamente que uma ação atingiu em um determinado período. Se o preço de fechamento é maior que o preço de abertura o *candlestick* é azul, caso contrário vermelho. Na figura 3.2 pode ser visualizado um *candlestick* e suas representações (máximo, mínimo, fechamento e abertura).

⁹ Pregão – seção durante a qual são negociadas as ações e títulos de uma empresa registrados em uma bolsa de valores. O pregão pode ser físico (diretamente na sala de negociações) e/ou eletrônico (pelo sistema de negociação eletrônica da bolsa de valores). Na Bmfovespa o horário de funcionamento padrão do pregão é das 10h às 17h. Sendo que no horário brasileiro de verão, o pregão regular funciona das 11h às 18h.

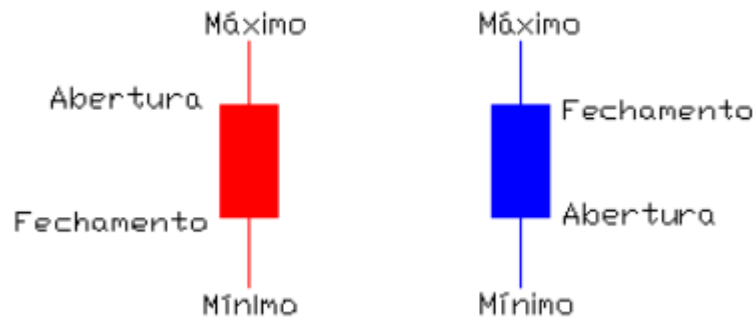


Figura 3.2 *Candlestick* e suas representações

A figura 3.3 abaixo representa um gráfico de *candlestick* correspondente à série temporal dos valores diários do Ibovespa e um indicador técnico (média móvel)¹⁰.

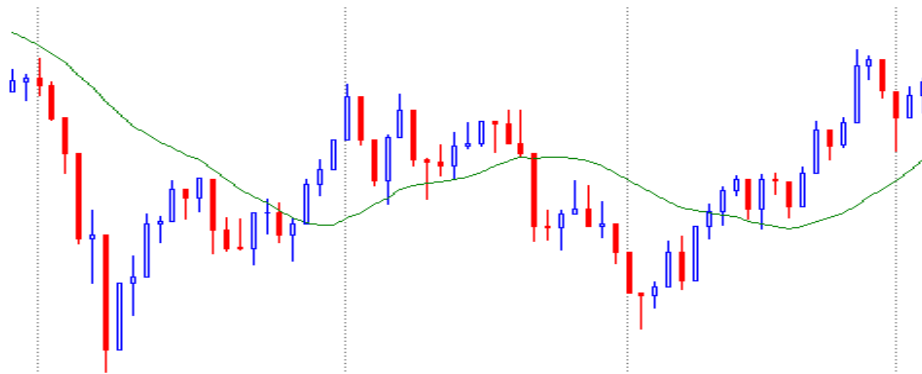


Figura 3.3 Gráfico de *candlestick*: Ibovespa (Período Agosto a Novembro de 2011)

É importante salientar a importância do Ibovespa, porque o mesmo reflete de uma maneira geral o comportamento do mercado como um todo. Logo se a tendência do indicador pode ser prevista fica mais fácil prever a tendência das principais ações do mercado, ou “*blue chips*”. Por isto a escolha do Ibovespa como proposta de estudo nesta dissertação.

¹⁰ Média móvel é um dos indicadores de tendência mais simples e popular utilizados na análise técnica. Pode ser considerado na prática como a média das últimas cotações do ativo para um determinado período de tempo escolhido para análise.

3.2 Análises de Investimento

Conforme citado anteriormente o valor de uma ação varia o tempo todo. Isto faz com que o investimento em ações seja de alto risco. Em investimento de risco, a liquidez de cada ação no mercado acionário está intimamente ligada ao momento presente da companhia e ao cenário global. As aquisições, vendas e fusões de companhias são as principais influências nos preços das ações, sendo também comum a oscilação de preço ser devido a fenômenos climáticos ou geopolíticos, por exemplo, empresas petrolíferas sofrem certa volatilidade com conflitos geopolíticos, etc. Devido a todos estes fatores citados acima descobrir quais ações comprar ou vender não é uma tarefa trivial. Neste sentido diferentes estudos e teorias foram desenvolvidos para tentar explicar o comportamento dos mercados de ações auxiliando assim os investidores na sua tomada de decisão. As duas teorias mais importantes (Hipótese do Mercado Eficiente e Teoria *Random Walk*) para previsão do mercado financeiro são apresentadas na próxima seção, assim como as análises técnicas e fundamentalistas que são duas abordagens convencionais que emergiram a partir das duas teorias principais.

3.2.1 Hipótese do Mercado Eficiente

A hipótese do mercado eficiente (*EMH*) foi introduzida por (Fama, 1964). A *EMH* diz que em um mercado eficiente o preço do mercado atual reproduz toda a informação disponível, ou seja, que as cotações dos ativos refletem toda a informação conhecida. Sendo assim o passado não contém qualquer informação que já não esteja incorporada no preço atual. Os preços variam com a chegada de novas informações.

Fama (1970) propôs três formas de eficiência de mercado (fraca, semi-forte e forte). A primeira delas (forma fraca) mostra que somente preços passados e informação histórica estão incorporados no preço corrente. Na segunda forma (semi-forte) os preços incorporam o seu comportamento passado e também o restante da informação publicada. Estas informações podem ser notícias específicas ou anúncios sobre distribuição de lucros e dividendos. Por fim, a forma forte, na qual os preços refletem não só a informação pública, mas toda a informação que pode ser obtida, inclusive as informações consideradas ocultas ou privilegiadas.

3.2.2 *Random Walk*

A teoria *Random Walk* (*RW*) é uma perspectiva diferente de previsão de preços no mercado financeiro. Nesta teoria a previsão do mercado de ações é considerada impossível, pois os preços são determinados de forma aleatória e superar o mercado é inviável. A teoria *RW* possui fundamentos teóricos similares à forma semi-forte da hipótese do mercado eficiente, onde toda a informação pública está disponível para todos. Entretanto a teoria *RW* afirma que mesmo com tais informações, a previsão futura é ineficiente (Schumarker *et al.*, 2009).

3.2.3 Análise Fundamentalista

A análise fundamentalista investiga os fatores que afetam a oferta e a procura das ações. O objetivo é reunir e interpretar estas informações e agir antes que a informação seja incorporada no preço das ações. O intervalo de tempo entre um evento e sua resposta de mercado apresenta uma oportunidade de negociação. A análise fundamentalista enfatiza principalmente os indicadores de desempenho de uma empresa, sua estratégia de negócios, mercado em que se insere a política de dividendos, política administrativa, solidez financeira, etc. Pode-se dizer então que neste tipo de análise são levadas em consideração informações como dados financeiros da empresa, dados macroeconômicos, planos e as estimativas da empresa para o setor, assim como a qualidade dos administradores, a situação interna do país e a sua posição em relação ao exterior. Neste sentido é possível afirmar que as notícias financeiras têm uma importância muito grande para investidores que utilizam análise fundamentalista, pois as mesmas descrevem os fatores que podem afetar o preço do ativo.

3.2.4 Análise Técnica

A análise técnica é baseada em dados de séries temporais numéricas e tenta prever o mercado de ações utilizando indicadores de análise técnica, construídos a partir do preço e outros parâmetros de negociação como volume, etc. Dentre estes indicadores é possível citar o índice de força relativa (*IFR*) das variações do preço das ações, médias móveis do preço ou do volume (*OnBalance Volume*), entre outros. A ideia é utilizar estes indicadores para a tomada de decisão na hora de investir, ou seja, determinar

tendências, pontos de entrada ou saída no mercado (*bullish*), entre outras. Nesta análise padrões de comportamento como triângulos retângulos, padrões ombro-cabeça-ombro e outros que ajudam a determinar o comportamento futuro da bolsa são identificados. A principal preocupação na análise técnica está relacionada na identificação de tendências existentes e assim tentar antecipar as tendências futuras do mercado acionário a partir de gráficos. Neste sentido notícias divulgadas no dia a dia não são utilizadas por especialistas que utilizam esta abordagem de previsão. Com base nesta abordagem e levando em consideração o histórico dos preços das ações diversos métodos de previsão podem ser implementados, como por exemplo, no trabalho de (Faria *et al.*, 2009), onde o valor do Ibovespa foi estudado e previsto através de métodos de redes neurais artificiais e alisamento exponencial adaptativo. Resultados neste trabalho reportaram que o indicador analisado pôde ser previsto com um percentual de acerto de 60%.

3.3 Análise Técnica x Análise Fundamentalista

Depois de tudo o que foi dito fica difícil argumentar por uma abordagem ou outra. Esta questão divide os investidores. Ao que tudo indica são os fatos associados a eventos fundamentalistas que conseguem reverter tendências ou desencadear movimentos bruscos do mercado. Já a análise técnica serve como ferramenta importante capaz de avaliar a força destes movimentos, e mais importante ainda serve para o investidor separar ou afastar o lado emocional na hora de tomar decisões, ou seja: substituir ideias como, “acho que vai subir” por “se romper nível de suporte a ação sobe”. É importante salientar que mesmo sem justificativa teórica a análise técnica deve ser acompanhada, uma vez que um grande número de investidores segue este tipo de decisões. E a auto-organização destes investidores em torno de um indicador técnico pode tornar válido esta abordagem mesmo sem nenhuma justificativa econômica.

4. Mineração de Textos para Análise de Notícias

O principal objetivo da Mineração de Textos (MT) é extrair informações relevantes em grande quantidade de dados não estruturados (textos). Weiss *et al.* (2005) afirmam que apesar dos dados utilizados na área de MT (textos) serem diferentes dos dados utilizados em Mineração de dados (numéricos), ambas apresentam características semelhantes, ou seja, são baseadas em amostras de exemplos passados e fazem uso dos mesmos métodos de aprendizagem. Sendo esta última característica justificada pelo fato dos textos serem processados e transformados em uma representação numérica similar a utilizada na Mineração de dados (MD). Para (Witten *et al.*, 2011) as áreas de MT e MD podem ser diferenciadas a partir de dois conceitos: A MD pode ser caracterizada como extração da informação implícita, anteriormente desconhecida, porém potencialmente útil a partir dos dados, ou seja, a informação está implícita nos dados de entrada e dificilmente poderia ser extraída sem recorrer às técnicas automáticas de MD. Já na MT a informação a ser extraída é clara, está explícita nos textos, no entanto o problema é que a informação não é expressa de uma maneira que é passível de processamento automático.

O fato é que com o crescimento de informação na forma não estruturada (textos), a MT ganha cada vez mais importância não só no meio acadêmico, mas também no mundo dos negócios.

De uma maneira geral a área de MT compreende a cinco tarefas: pré-processamento de textos, sumarização, classificação automática de textos, agrupamento de textos e recuperação de informação.

Conforme citado na introdução deste trabalho, quando se está trabalhando com MT e mercado financeiro existem diferentes estratégias (estudo de eventos, classificação de textos e análise de sentimento) para diferentes objetivos. Como o principal objetivo deste trabalho é prever a tendência diária do comportamento futuro do índice do mercado acionário do Brasil utilizando o título de notícias, pode-se dizer então que este objetivo está dentro do contexto da estratégia de classificação de textos. Sendo assim ênfase é dada apenas à tarefa de classificação automática de textos e consequentemente a tarefa de pré-processamento dos textos, uma vez que a mesma é o passo preliminar requerido não só a tarefa de classificação de textos, mas também a

qualquer uma das tarefas de MT citadas anteriormente. Detalhes sobre as demais tarefas podem ser encontradas em (EBECKEN *et al.*, 2003).

4.1. Classificação de Textos

A classificação automática de textos (CT) tem assistido a um crescente interesse nos últimos anos devido ao crescimento e a disponibilidade de documentos online. A abordagem dominante para este problema na comunidade científica é baseada em técnicas de aprendizado de máquina, ou seja, um processo geral indutivo constrói automaticamente um classificador por “aprendizado”, a partir de um conjunto de documentos classificados previamente e características de uma ou mais categorias. A vantagem desta abordagem sobre a que era utilizada anteriormente (Engenharia do conhecimento)¹¹ é o fato de ter uma melhor eficácia, uma economia considerável em termos de mão de obra especializada e independência de domínio (Sebastiani, 2002).

Pode-se dizer então que a classificação automática de textos é uma tarefa de aprendizagem de máquina com o problema de atribuir automaticamente categorias pré-definidas a documentos textuais.

Formalmente Sebastiani (2002) afirma que a CT consiste em determinar se um documento d_i (de um conjunto de documentos D) pertence ou não a uma categoria c_j (de um conjunto de categorias C), consistentemente com o conhecimento das categorias corretas para um conjunto de documentos de treinamento.

4.1.1 Aplicações da Classificação de Textos

A classificação automática de textos está se tornando cada vez mais importante devido à facilidade de lidar com grandes quantidades de dados em massa tais como bibliotecas digitais, fontes de notícias e informações internas das empresas. Com o advento da internet a classificação de textos tem permitido a organização de um grande volume de informações que são lançadas na *web* todos os dias. Com isto novos métodos

¹¹ Na Engenharia do conhecimento regras são definidas manualmente por um engenheiro do conhecimento com o auxílio de um especialista no domínio da aplicação. Estas regras definem como classificar documentos a partir de características dos textos.

baseados em teorias estatísticas e aprendizado de máquina têm sido aplicados na categorização de textos nos últimos anos.

Sebastiani (1999) afirma que a classificação de textos é atrativa porque liberta as empresas da necessidade de organizar grandes quantidades de documentos manualmente, o que pode ser muito caro, ou muitas das vezes um trabalho inviável dada às limitações de tempo da aplicação ou mesmo ao número de documentos envolvidos.

Nos últimos anos a classificação de textos vem sendo utilizadas em diferentes aplicações como, por exemplo: na organização de documentos, Filtragem de Textos (onde um filtro pode, por exemplo, determinar se um e-mail é ou não um spam), Categorização Hierárquica de Páginas da *Web*, entre outras. Detalhes sobre estas e outras aplicações podem ser encontradas em (Sebastiani, 2002).

Dumais *et al.* (1998) afirma que a classificação de textos pode desempenhar um papel importante em uma ampla variedade de tarefas de gestão de informação mais flexíveis e dinâmicas tais como: classificação em tempo real de e-mails ou arquivos em pastas hierárquicas, identificação de tópicos, pesquisa ou navegação estruturada e a busca por documentos baseados em interesses.

4.1.2 Processo Geral de Classificação Automática de Textos

A classificação de textos envolve vários passos e processos. Diferentes trabalhos existentes na literatura identificam estes processos de várias maneiras e com diferentes nomes.

Para (Apte *et al.*, 1994) o processo de classificação de textos envolve quatro principais processos: **Pré-processamento**: para determinar os valores das características, atributos ou palavras chaves que serão usados na representação individual do documento dentro de uma coleção. Este processo é essencialmente a criação do dicionário. **Representação**: para mapear cada documento individual em uma amostra de treinamento usando o dicionário criado no processo anterior e associando-o com um rótulo que identifica a sua categoria. **Indução**: para encontrar padrões que distingue uma categoria das outras. E por fim a **Avaliação**: para escolher a melhor solução com base em minimizar o erro ou custo do classificador.

Já para (Debole *et al.*, 2003) o processo de classificação de textos pode ser resumido em três fases: **Seleção de termos**: onde os termos mais relevantes para a tarefa de classificação são identificados. **Ponderação dos termos**: onde é realizado o

cálculo dos pesos para os termos selecionados. **Aprendizagem do classificador:** geração de um classificador a partir das representações ponderadas dos documentos de treinamento. Podendo os dois primeiros processos (seleção e ponderação dos termos) serem identificados como uma única fase chamada de **Indexação de documentos**, o que significa a criação de representações internas dos documentos.

Bastos (2006) declara que o processo de classificação de textos pode ser implementado em quatro etapas: **Obtenção de documentos:** onde os textos relacionados à aplicação são obtidos. **Extração de características ou Pré-processamento:** preparação dos textos para o processo de mineração. Nesta etapa podem ser utilizados métodos para retirada de símbolos, pontuação e eliminação das *stop words* (palavras que não agregam valor por não possuir significado específico); métodos para identificação de termos dos textos (processo de *stemming* e técnicas de “*n-gramas*”); métodos para redução de termos e métodos para a definição de pesos para os termos (frequência relativa ou frequência inversa). **Extração de Conhecimento:** utiliza algoritmos de aprendizagem com o objetivo de extrair, a partir dos textos pré-processados, conhecimento na forma de regras de associação, relações, segmentação, classificação de textos, entre outros. E por fim a última etapa, que está associada à **Avaliação e Interpretação dos resultados:** os dados obtidos são analisados, filtrados e selecionados.

Neste trabalho o processo de CT considerado é constituído de quatro etapas principais conforme visualizado na figura 4.1 abaixo. Uma explicação detalhada de todas estas etapas é apresentada nas próximas seções.

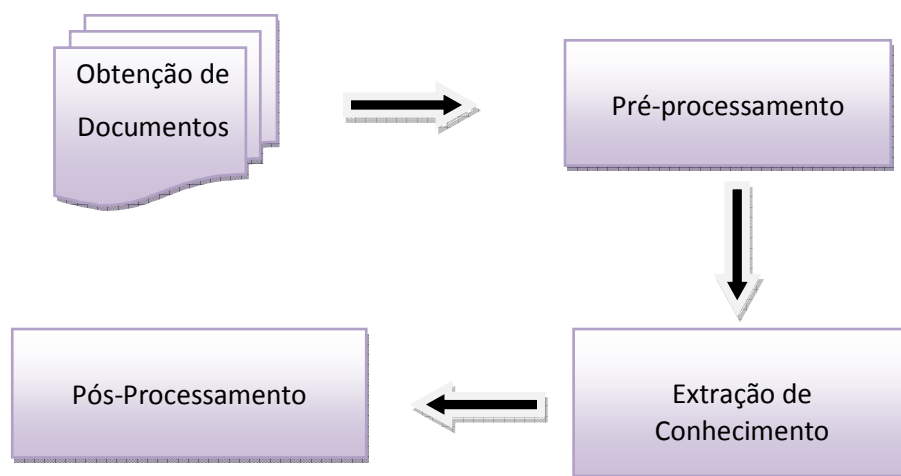


Figura 4.1 Arquitetura geral dos processos de Classificação de textos

A primeira etapa a ser realizada é a Obtenção dos documentos, cujo objetivo é a formação da coleção de documentos que serão utilizados para a tarefa de classificação. Em seguida, inicia-se a etapa de pré-processamento. Nesta etapa os documentos obtidos anteriormente são submetidos a várias operações, a fim de atingir uma estruturação que permita a extração de conhecimento. A extração de conhecimento é a etapa que faz uso de algoritmos de aprendizagem de máquina para extrair conhecimento a partir dos textos pré-processados, para a classificação dos textos. E finalizando o processo de CT temos a etapa de pós-processamento que visa avaliar e interpretar os resultados obtidos.

4.1.2.1 Obtenção de Documentos

Este é o primeiro processo na classificação de textos. O principal trabalho neste processo é a busca por documentos (textos) que são relevantes para a aplicação. No caso específico deste trabalho estamos interessados em notícias financeiras, ou seja, notícias que podem afetar o comportamento do mercado da bolsa de valores do Brasil. Sendo assim as fontes destes dados são sites de notícias econômicas e a obtenção dos mesmos é feita de forma automática.

É importante salientar que existem diferentes fluxos de informações e notícias que podem afetar o comportamento dos mercados financeiros. Neste sentido tão importante quanto à obtenção dos dados é distinguir quais são os tipos de notícias que são relevantes quando se está tentando entender o comportamento do mercado financeiro, em especial o mercado de ações. Leinweber (2009) classifica os diferentes tipos de notícias disponíveis na *web* em quatro categorias:

- 1 Notícias - Compreende a grandes mídias e notícias produzidas por fontes confiáveis. São geralmente transmitidas através de jornais, rádios e televisão. Elas também são distribuídas para as mesas dos investidores através de agências de notícias. A grande maioria possui suas versões online também.

- 2 Pré-notícias – Fontes de dados que são investigadas antes de serem publicadas. Geralmente são relatórios da *Securities and Exchange Commission*¹² e documentos judiciais e governamentais. Inclui também anúncios regulares tais como notícias macroeconômicas, estatísticas industriais, relatórios e lucros das empresas e outras notícias corporativas.
- 3 Boatos – *Blogs* e *websites* que transmitem notícias e são menos respeitáveis do que fontes de notícias e pré-notícias. A qualidade destas notícias varia significativamente. Alguns *blogs* podem estar associados com provedores de notícias e respeitáveis repórteres (por exemplo, o *blog* de *Robert Preston* da *BBC*). Outros podem ser abastecidos totalmente por boatos.
- 4 Mídia Social – *Websites* onde não existem impedimentos para publicação de notícias, ou seja, não há uma fiscalização adequada que monitore as mensagens publicadas. Estes podem ser fontes de informações perigosamente imprecisas. No entanto se aplicada com muito cuidado pode haver algum valor. No mínimo pode ser utilizado para ajudar a identificar volatilidade futura.

Alguns estudos vêm sugerindo que diferentes tipos de investidores prestam atenção em diferentes tipos de notícias, como o trabalho de (Dzielinski *et al.*, 2010), onde os autores concluíram que investidores individuais prestam mais atenção nas fontes 1 e 2, ou seja notícias e pré-notícias respectivamente. Já Leinweber (2009) afirma que mesmo que as notícias disponíveis na *web* sejam menos confiáveis, se um grupo grande de pessoas participa de uma mesma opinião (sem segundas intenções), esta opinião pode ser útil para os investidores.

¹² A *Securities and Exchange Commission* (SEC) é uma agência federal dos Estados Unidos que detém a responsabilidade primária pela aplicação das leis de títulos federais e a regulação do setor de valores mobiliários, as ações da nação e opções de câmbio e outros mercados de valores eletrônicos nos Estados Unidos. No Brasil a responsabilidade de regular o mercado financeiro é da CVM (Comissão de Valores Imobiliários).

4.1.2.2 Pré-processamento de Notícias

O pré-processamento é o segundo processo na arquitetura de CT. É um dos principais processos, pois resultados de um pré-processamento feito erroneamente pode comprometer os próximos processos da CT.

Wang *et al.* (2005) consideram o pré-processamento como um processo de tornar claro cada estrutura da linguagem e eliminar, tanto quanto possível os fatores dependentes do idioma. Lee e Chen (2006) definem esta tarefa apenas como remoção de *stop-words* e *stemming*. Para (Wei *et al.*, 2001) a fase de pré-processamento consiste em transformar os dados textuais obtidos na etapa anterior em uma representação adequada a ser utilizada pelos algoritmos de aprendizagem e compreende a três atividades, extração de características, seleção de características e representação do documento. Devido ao fato desta abordagem ser uma das mais utilizadas na literatura optou-se pela descrição de pré-processamento dentro do contexto proposto por estes autores. Sendo assim estas atividades são descritas na próxima seção.

Extração de Características

A extração de características é o primeiro passo no pré-processamento de documentos. O Objetivo principal é gerar uma lista de termos que descreve suficientemente os documentos. Portanto os documentos de treinamento são analisados para determinar uma lista de todas as características (palavras-chave) contidas nos documentos. Em seguida técnicas de redução de características são aplicadas para reduzir a dimensão da lista criada (frequentemente indicada como dicionário). Os métodos mais utilizados para esta finalidade são a eliminação de *stop words*, utilização de algoritmos de *stemming* e utilização de um dicionário ou *thesaurus*.

As *stop words* conforme citado anteriormente são palavras que podem ser eliminadas por não constituírem conhecimento nos textos. Geralmente são dadas como uma *stop list* (lista com as palavras a serem descartadas). A maioria destas palavras são preposições, pronomes, artigos e outras classes de palavras auxiliares, que não contribuem para o processo de classificação e muitas das vezes têm influência negativa. O principal objetivo da eliminação das *stop words* é reduzir o ruído presente no dicionário criado.

O processo de *stemming* consiste em reduzir cada palavra do texto em sua provável palavra raiz. O objetivo é remover sufixos e prefixos dos termos que possam vir a identificar plurais, gênero ou formas verbais, fazendo com que assim termos equivalentes tenham uma única representação. Este processo reduz o número de termos diferentes nas representações dos documentos e impede também a repetição de palavras com mesmo significado conceitual (mesmo radical). Os algoritmos de *stemming* fazem uso de regras linguísticas e são dependentes do idioma. Em se tratando da língua portuguesa, o algoritmo de *stemming* mais utilizado nas ferramentas é o *Portuguese Stemming* proposto por (Orengo *et al.*, 2001). Este algoritmo foi implementado em C e é composto de várias etapas. Entretanto por não fazer parte do escopo deste trabalho não daremos ênfase a implementação deste algoritmo. Detalhes sobre o mesmo podem ser encontrados em (Bastos, 2006).

Resultados experimentais publicados em (Baker *et al.*, 1998) demonstraram que em muitos casos o processo de *stemming* pode ser prejudicial à eficácia do classificador de textos. Entretanto vale ressaltar que apoiadores do método afirmam que o mesmo é importante sim, não só pelo fato de reduzir a dimensionalidade do espaço de características, mas também é útil para ajudar a promover a eficácia do classificador de textos.

Um *thesaurus* ou dicionário é um vocabulário controlado que representa as palavras que tem o mesmo valor semântico, ou seja, termos diferentes existentes nos documentos com o mesmo valor semântico podem ser substituídos por termos-chave com o auxílio de um dicionário de sinônimos. O caso mais simples de *thesaurus* é um par de sinônimos, isto é, duas palavras que tem o mesmo significado, como por exemplo, as palavras carro e automóvel, que poderiam estar associadas a uma única palavra que é “carro”. A maior dificuldade para a utilização de um *thesaurus* é a criação do próprio dicionário.

Seleção de Características

O principal objetivo desta fase é eliminar aquelas características que fornecem pouca ou nenhuma informação importante e assim diminuir a dimensionalidade do espaço de características. Shang *et al.* (2006) afirma que o maior problema na tarefa de classificação de textos é a alta dimensionalidade do espaço de características e que para

muitos algoritmos de aprendizagem esta alta dimensionalidade não é adequada. Montanes *et al.* (2003) afirma que uma grande quantidade de características utilizadas para representar os documentos na CT é irrelevante e as mesmas podem ser eliminadas. Para (Forman, 2003) uma boa seleção de características é essencial para melhorar o desempenho do classificador.

Pode-se dizer então que a seleção de características é uma forma de reduzir a alta dimensionalidade do espaço de características selecionando apenas as características mais importantes dos textos.

Existem duas abordagens para a seleção de características em CT, a abordagem *Filter* e abordagem *Wrapper* (Sebastiani, 1999). A abordagem *Filter* opera independente do algoritmo de classificação considerado e seleciona as características antes do processo de classificação e geralmente com base em medidas estatísticas, enquanto a abordagem *Wrapper* requer um algoritmo de aprendizado para selecionar as características e utiliza o seu desempenho para avaliar e determinar quais características devem ser selecionadas, ou seja, para cada subconjunto de características considerado é avaliado diretamente o desempenho do classificador. As figuras 4.2 e 4.3 apresentam estas duas abordagens.

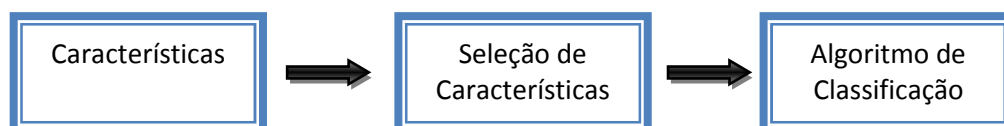


Figura 4.2 Esquema de abordagem *Filter*

Características selecionadas independente do algoritmo de classificação.

Adaptado de (John *et al.*, 1994).

A utilização da abordagem *Filter* pode ser relativamente rápida, porém os resultados podem não ser satisfatórios quando aplicados ao algoritmo de classificação. John *et al.* (1994) afirma que isto ocorre devido ao fato do algoritmo de classificação ser ignorado no processo de seleção de características. Chou *et al.* (2010) declara que a abordagem *Wrapper* pode encontrar o melhor subconjunto de características para um determinado algoritmo de aprendizagem, entretanto o custo computacional é muito maior e pode tornar impraticável quando o número de características originais for muito grande. Segundo (Doan *et al.*, 2004) o método *Wrapper* é relativamente difícil de ser implementado especialmente quando a quantidade de dados é muito elevada. Por outro

lado a abordagem *Filter* é facilmente entendida e computacionalmente eficiente, o que justifica o uso frequente da mesma em aplicações envolvendo CT. Chou *et al.* (2010) também afirma que a abordagem *Filter* é mais popular e comumente utilizada.

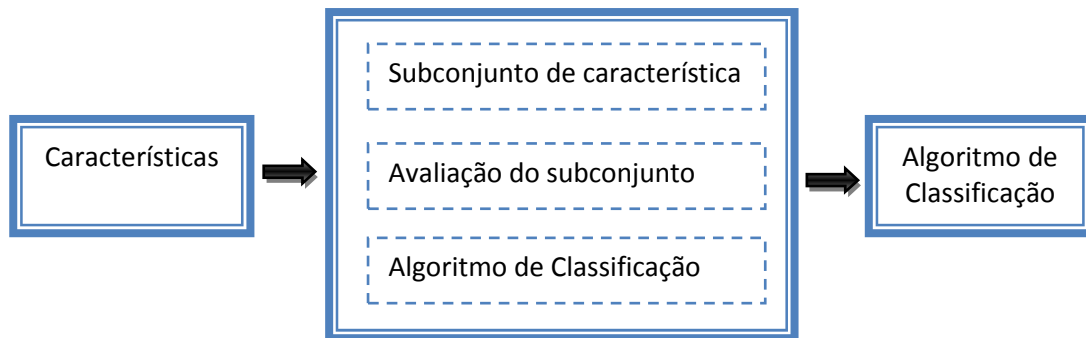


Figura 4.3 Esquema de abordagem *Wrapper*

Utilização do algoritmo de classificação na fase de seleção de características.

Adaptado de (John *et al.*, 1994).

A abordagem utilizada neste trabalho é a mais comumente utilizada que é a abordagem *Filter*. Conforme citado anteriormente a abordagem *Filter* faz uso de métricas estatísticas para selecionar as características mais relevantes dos documentos. Existem na literatura diversas medidas estatísticas para esta finalidade, como por exemplo, Intensidade do Termo, Coeficiente de Correlação, Frequência do documento, Ganho de Informação, Estatística X^2 , Informação Mútua, *etc.*, entretanto as mais comumente utilizadas são:

- ✓ Frequência de Documentos – *Document Frequency* (DF): frequência de documentos é o número de documentos no qual um termo aparece. A DF para cada termo no conjunto de treinamento é computada e os termos cuja frequência de documentos é inferior a um limite predeterminado são removidos do espaço de características. A ideia básica desta medida é o fato de que se o termo não aparece em muitos documentos então ele não pode ser considerado como um bom indicador das classes e conseqüentemente o mesmo pode ser removido reduzindo assim a dimensionalidade do espaço de características. Este método é utilizado em (Apte *et al.*, 1994) e também em (Yang *et al.*, 1997) tendo este último concluído que é possível fazer uma grande redução de características sem perda de eficiência do classificador utilizado.

- ✓ Ganho de Informação – *Information Gain* (IG): Ganho de Informação é um método baseada na entropia. É frequentemente empregado como um critério de importância do termo no campo do aprendizado de máquina (Mitchell, 1998). Ele mede o número de partes de informação obtidas para predição da categoria, conhecendo-se a presença ou ausência de um termo em um documento. Para cada termo é calculado o ganho de informação e os termos cujos ganhos de informação forem menores que um limiar determinado podem ser removidos do espaço de características. Este cálculo inclui as estimativas das probabilidades condicionais de uma categoria dado um termo e o cálculo da entropia na definição. Lewis e Ringuette (1994) utilizaram o Ganho de Informação para a seleção de características em um modelo para classificação binária, juntamente com os algoritmos *Naive Bayes* e *Árvore de decisão*.

- ✓ Informação Mútua – *Mutual Information* (MI): A informação mútua é um critério normalmente utilizado em modelagem estatística da linguagem em associações de palavras e aplicações correlatas (Church e Hanks, 1990). Dumais *et al.* (1998) utilizaram MI para seleção de características juntamente com diferentes algoritmos de aprendizagem. Uma deficiência desta métrica é que a pontuação é fortemente influenciada pelas probabilidades marginais dos termos. Para termos com uma probabilidade igual a $P_r(t/c)$ (probabilidade do termo t acontecer, dado que a classe é c), poucos termos vão ter uma pontuação maior que termos comuns. Desta forma, não se pode comparar pontuações entre termos de frequência muito diferentes.

- ✓ Estatística χ^2 (CHI) - A estatística CHI mede a falta de independência entre cada termo e as categorias e pode ser comparada a distribuição CHI com um grau de liberdade para julgar extremos. Quando o valor da estatística CHI é zero significa que termo e categorias são independentes. Sendo assim apenas aqueles termos que possui a estatística CHI maior que um limiar predeterminado são selecionados e os demais podem ser descartados, pois uma vez que os mesmos são independentes à categoria eles não possuem nenhum significado para a classificação. A maior diferença entre a estatística CHI e a informação mútua é que o CHI é um valor normalizado e, portanto seu valor é comparável entre os

termos de uma mesma categoria. No entanto esta normalização é ruim para termos de baixa frequência. Logo a estatística CHI não é considerada muito confiável para termos de baixa frequência (Dunning, 1993).

Mais informações sobre as métricas citadas, assim como a formulação matemática das mesmas podem ser encontradas em (Sebastiani, 2002). E muitos trabalhos visando o problema de seleção de características podem ser encontrados na literatura. Como por exemplo, no trabalho de (Yan *et al.*, 1997) e apoiados por (Sebastiani, 2002) os autores sugerem que para diferentes classificadores e diferentes coleções de documentos os métodos CHI e IG têm similar desempenho e podem reduzir a dimensionalidade do espaço de características (redução de 98%) sem perder a eficácia e que os resultados utilizando DF (redução de 90%) também podem ser comparáveis aos desempenhos do CHI e IG. Entretanto os resultados da MI (redução de 50-60%) foram inferiores aos métodos anteriores. Segundo os autores a MI teve resultados inferiores devido a uma tendência de favorecer termos raros e uma forte sensibilidade a erros de estimação das probabilidades.

Cai e Song (2008) introduziram um novo método de seleção de características chamado “*Count Difference*” que é baseado na diferença entre as frequências relativas do documento (relação de frequência do documento de uma característica para uma classe sobre a frequência média do documento para a mesma classe) de uma característica para classes relevantes e irrelevantes. Os autores compararam o método proposto com outros métodos de seleção de características (DF, MI, IG, entre outros). Estes métodos de seleção de características foram avaliados em uma base de dados de classificação de textos (*Reuters*) com a modelagem da máxima entropia. Por fim os autores mostraram que a seleção de característica é uma forma de redução do custo computacional e ao mesmo tempo melhora o desempenho do classificador. Eles demonstraram também que o método de seleção proposto (*Count difference*) é promissor para classificação de textos, não somente por obter melhor desempenho, mas também por trabalhar razoavelmente bem com cortes de características muito agressivos. E finalmente quanto ao classificador os autores demonstraram que a modelagem de máxima entropia é um método competitivo para classificação de textos quando comparado a outros algoritmos de classificação (SVM, k - NN e *Rocchio*).

Apesar desta vasta pesquisa em relação a diferentes métodos de seleção de características, é possível concluir que não existe uma clara evidência da superioridade de qualquer método em particular para todos os tipos de dados, ou seja, cada método apresenta suas qualidades e fraquezas. Como os métodos de seleção de características são dependentes dos dados e dos algoritmos de classificação utilizados fica difícil dizer qual método é mais eficiente.

Representação do documento

A representação do documento é a tarefa final no pré-processamento do documento. Os documentos são representados em termos daquelas características para os quais o dicionário foi reduzido nas etapas anteriores. Assim, a representação de um documento é um vetor característica de n elementos, onde n é o número de características remanescente após o término do processo de seleção.

Para (Strzalkowski, 1994) a representação do documento tem a finalidade de reduzir a complexidade dos documentos textuais e torná-los mais fáceis de serem trabalhados. Os documentos têm que ser transformados a partir de sua versão textual original para um “vetor documento” que descreve o seu conteúdo. Um documento pode ser representado como uma coleção de características: palavras, frases, sentenças, entre outros, derivadas a partir dos documentos textuais. Estas características são geralmente ponderadas para indicar sua importância.

A abordagem universal mais simples para a representação de documentos é a representação “saco de palavras” (do inglês, *bag of words*) também conhecida como Modelo de Espaço Vetorial. O Modelo de Espaço Vetorial segundo (Salton *et al.*, 1975) é um modelo algébrico para representação de documentos textuais como vetores de características que identificam cada um dos documentos no modelo. Os documentos são representados por vetores e cada dimensão no vetor documento corresponde a uma única característica, a qual possui um peso associado.

O principal problema da utilização deste modelo está na alta dimensionalidade inerente a Mineração de Textos, pois dado um *corpus* com algumas centenas de documentos, o número de características facilmente pode ultrapassar centena de milhares muito facilmente.

Nesta representação a ordem e a ligação entre as características não tem nenhum valor para o sistema. Pode se dizer então que esta abordagem é visivelmente pobre em relação a todos os recursos que o vocabulário de uma língua pode oferecer, inviabilizando muitas técnicas de Processamento de Linguagem Natural. No entanto, a codificação *bag of words* tem apresentado bons resultados na literatura, justificando assim a sua abordagem puramente estatística. (Boulis *et al.*, 2005) afirma que apesar da simplicidade da representação *bag of words*, os métodos de classificação que fazem uso da mesma, muitas vezes apresentam um alto desempenho. Na figura 4.4 abaixo pode ser visualizado a representação de um documento utilizando o modelo *bag of words*.

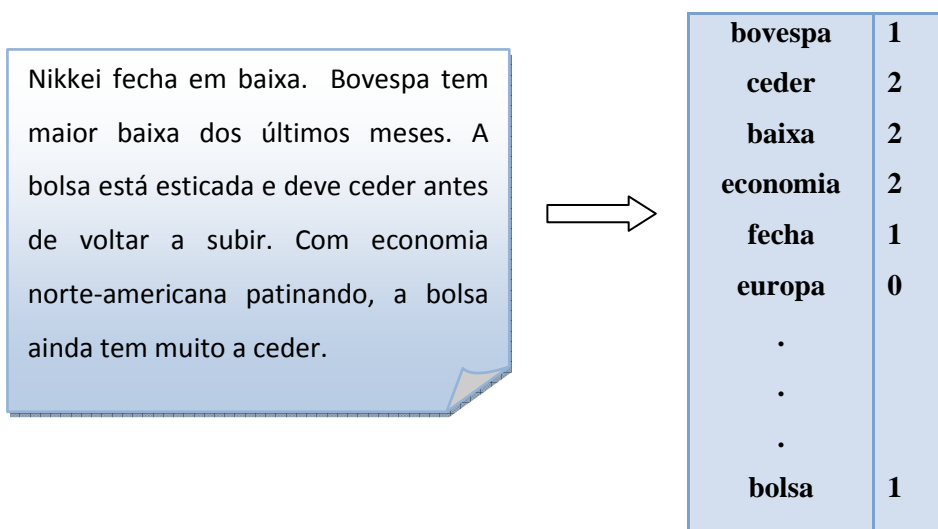


Figura 4.4 Representação *Bag of Words*

Conforme pode ser observado na figura acima, os termos apresentados estão associados a seus respectivos pesos. De uma maneira geral a atribuição destes pesos pode ser feita através de vários métodos, dentre os quais podemos citar:

- ✓ Medida Booleana – usa representação binária para os termos. Se o termo t_j está presente no documento d_i o valor do peso a_{ij} é 1, caso contrário é zero. Por ser muito simples, esta medida de atribuição de pesos é raramente utilizada. Salton e Buckley (2002) declara que medidas estatísticas levando em consideração a frequência dos termos nos documentos geralmente são as mais empregadas.

- ✓ TF (*term frequency*) – cada termo é assumido ter importância proporcional ao número de vezes que ele ocorre no documento. Sendo assim o peso a_{ij} de um termo t_j em um documento d_i pode ser dada pela seguinte fórmula:

$$tf(t_j, d_i) = freq(t_j, d_i) \quad (4.1)$$

Onde $freq(t_j, d_i)$ é frequência do termo t_j no documento d_i . A TF nem sempre é uma boa medida de atribuição de pesos, porque o fato de termos frequentes aparecerem na maioria dos documentos significa que eles tem pouco poder de discriminação. Uma boa solução para este problema é a remoção dos termos com alta frequência do conjunto de características. Entretanto encontrar um limiar ideal é o grande problema desta abordagem (Tokunaga *et al.*, 1994). Peramunetilleke *et al.*, (2001) também afirma que o TF sozinho não é um bom indicador para a seleção de características preditivas, pois o termo que aparece frequentemente não é necessariamente um indicador para avaliar a subida ou descida do valor de uma ação.

- ✓ IDF (*inverse document frequency*) – diferentemente da TF (que se preocupa com a ocorrência do termo dentro de um texto) a IDF se preocupa com a ocorrência do termo em uma coleção de textos. A ideia básica da IDF é que termos que ocorrem raramente em uma coleção de textos são de grande valor. Assume-se que a importância de cada termo é inversamente proporcional ao número de textos que contém o termo. A IDF de cada termo t_j pode ser calculada por:

$$idf(t_j) = \log \frac{N}{d(t_j)} \quad (4.2)$$

Onde N é o número total de textos na coleção e o denominador é o número de textos que contém o termo t_j pelo menos uma vez.

- ✓ TF-IDF (*term frequency-inverse document frequency*) - a TF-IDF é uma medida bastante citada na literatura. Proposta por (Salton *et al.*, 1975) a TF-IDF é o produto da combinação da TF e da IDF (ambas citadas anteriormente) para a ponderação de termos e pode ser calculada pela seguinte fórmula:

$$tf - idf (t_j, d_i) = freq (t_j, d_i).idf (t_j) \quad (4.3)$$

Através da medida *tf-idf* é possível aumentar a importância dos termos que aparecem em poucos documentos e diminuir a importância de termos que aparecem em muitos documentos. Salton e Buckley (2002) e Sebastiani (2002) afirmam que esta medida é a mais utilizada em aplicações envolvendo classificação de textos.

4.1.2.3 Extração de Conhecimento

Na fase de extração de conhecimento um classificador é construído a partir das representações ponderadas do conjunto de treinamento (previamente classificado) obtidas após a fase de representação do documento. Sendo assim o classificador construído é capaz de aprender os padrões necessários que distinguem uma classe da outra.

As estratégias de aprendizagem para aprender automaticamente os padrões em classificação de textos podem ser divididas em diferentes tipos: árvores de decisão (Lewis e Ringuette, 1994), redes neurais artificiais (Chen *et al.*, 1994), (Yang, 1994), *K-nearest neighbor* (Kwon e Lee, 2000), Naive Bayes (Lewis e Ringuette, 1994), algoritmo de *Rocchio* (Joachims, 1997), *Support Vector Machines* (Joachims, 2003), e métodos baseados em regressão (Schutze *et al.*, 1995).

A seguir são apresentados os algoritmos de aprendizagem utilizados neste trabalho (*Naive Bayes*, Redes Neurais Artificiais e *Support Vector Machines*). Para mais detalhes sobre os mesmos e também dos demais métodos citados acima, ver (Sebastiani, 1999).

Naive Bayes

O classificador *Naive Bayes* (NB) é denominado ingênuo (*naive*) por assumir que existe uma independência entre os termos de um documento, ou seja, a informação de um evento não é informativa sobre nenhum outro. Segundo (Zhang, 2004) apesar de ser considerado “ingênuo” e simplista, o classificador é um dos mais utilizados em

Aprendizado de Máquina e na maioria das vezes apresenta bom desempenho em várias tarefas de classificação.

O *Naive Bayes* é baseado no teorema de *Bayes* cuja ideia principal é utilizar a junção de probabilidades das palavras e classes para estimar as probabilidades das classes de um novo documento. De um modo geral pode-se dizer então que o algoritmo através da regra de *Bayes* calcula a probabilidade a posteriori de um documento pertencer a classes diferentes e o atribui a classe cuja probabilidade a posteriori é a mais alta. Sendo assim a um novo documento é atribuído à classe com a mais alta probabilidade a posteriori.

As principais vantagens do classificador NB é a simplicidade do algoritmo, o que facilita a sua implementação e o fato de não necessitar de um procedimento de aprendizagem, pois as probabilidades calculadas são estimadas através das frequências dos termos.

A desvantagem do NB é o fato do mesmo assumir que os atributos são independentes dadas à classe, hipótese que raramente pode ser confirmada no mundo real.

Support Vector Machines

Support Vector Machines ((Vapnick, 1995), (Vapinick, 1998)) são um grupo de métodos de aprendizagem supervisionado que executa a classificação através da construção de um hiperplano ótimo N -dimensional que separa os dados em duas categorias. Nos últimos anos a SVM tem demonstrado um desempenho muito bom em uma ampla variedade de aplicações envolvendo classificação que exigem espaço de entrada em larga escala, tais como reconhecimento de caracteres manuscritos (LeCun *et al.*,1995), detecção de face (Osuna *et al.*,1997) e o mais importante neste caso, a classificação de textos ((Joachim, 1998); (Dumais *et al.*, 1998)).

A SVM é baseada na teoria de aprendizagem estatística, que faz uso do princípio da minimização do risco estrutural (*Structural Risk Minimization* - SRM) ao invés da minimização do risco empírico (*Empirical Risk Minimization* - ERM) que é comumente utilizado em outros métodos estatísticos, como é o caso das Redes Neurais Artificiais. A SRM está baseada no fato de que o erro do algoritmo de aprendizagem junto aos dados de validação (dados que tem a mesma distribuição dos dados de treinamento, porém ainda não apresentado ao algoritmo), ou seja, o erro de generalização é limitado pelo

erro de treinamento mais um termo que depende da dimensão VC (dimensão *Vapnik e Chervonenkis*), que é uma medida da capacidade de expressão de uma família de funções classificadoras obtidas por meio de um algoritmo de aprendizagem (Vapnik, 1995). O objetivo é construir um conjunto de hiperplanos cuja estratégia é a variação da dimensão VC, de forma que o risco empírico (erro de classificação calculada no conjunto de dados de treinamento) e a dimensão VC sejam minimizados ao mesmo tempo.

De uma maneira geral pode-se dizer que uma SVM pode ser descrita da seguinte forma: dadas duas classes e um conjunto de pontos que pertencem a essas classes, uma SVM determina o hiperplano que separa os pontos de tal forma que o maior número de pontos da mesma classe é colocado do mesmo lado, enquanto maximiza a distância de cada classe a esse hiperplano. A distância de uma classe a um hiperplano é a menor distância entre eles e os pontos dessa classe e é chamada de margem de separação. O hiperplano gerado pela SVM é determinado por um subconjunto dos pontos das duas classes, chamados de vetores de suporte.

Para Joachim (1998) a evidência teórica que torna a utilização das SVMs tão atrativas em aplicações envolvendo a classificação de textos é o potencial das mesmas para lidar com uma alta dimensionalidade do espaço de características, como é o caso das aplicações envolvendo classificação de textos.

Artificial Neural Network

A Rede Neural Artificial (RNA) (do inglês *Artificial Neural Network*) é uma forma de computação não algorítmica e caracterizada por sistemas que, em algum nível, relembram a estrutura do cérebro humano, sendo assim constitui uma alternativa à computação algorítmica convencional (Braga, 2000).

De uma maneira geral uma rede neural pode ser considerada como uma estrutura de processamento de informação paralelamente distribuída, constituída de duas estruturas principais: (1) os nós, que correspondem aos neurônios ou unidades de processamento; (2) os links, que correspondem às conexões sinápticas entre os neurônios e que estão associadas a um determinado peso. O valor destes pesos é determinado através de um processo de treinamento (ou aprendizado) das redes e a distribuição dos mesmos em essência armazena o conhecimento e as relações determinantes do sistema estudado. Os neurônios de uma rede neural são organizados

em camadas e o comportamento de uma camada depende da sua função de ativação e do padrão de conexão entre os pesos. A figura 4.5 exemplifica uma arquitetura básica das Redes Neurais e seus principais componentes.

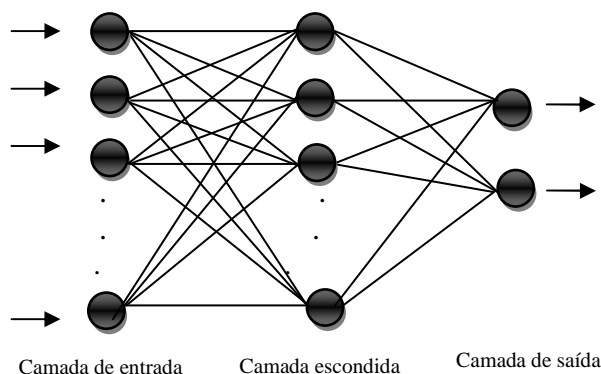


Figura 4.5 Arquitetura básica de uma rede neural e seus principais componentes

Uma das propriedades mais importantes das redes neurais e que capta bem a característica humana é a habilidade de aprender. Isto é feito através de um processo iterativo de ajustes aplicados a seus pesos, chamado treinamento. O objetivo do treinamento de uma rede neural é fazer com que a aplicação de um conjunto de entradas produza um conjunto de saídas desejado ou no mínimo um conjunto que apresente saídas consistentes, onde o erro seja mínimo.

Os diferentes modelos de RNAs podem ser diferenciados pelo seu tipo de treinamento que pode ser supervisionado ou não supervisionado e também pelo tipo de arquitetura (Redes Alimentadas Adiante com Camada Única, Redes Alimentadas Diretamente com Múltiplas Camadas e Redes Recorrentes) (Haykin, 2000).

Dentre os modelos mais citadas na literatura estão os *Perceptrons de Múltiplas Camadas* (do inglês *Multilayer Perceptron* - MLP), as redes de *Função de Base Radial* (do inglês *Radial Basis Function* - RBF) e as redes de *Kohonen*. Neste trabalho em específico foi abordado o problema de classificação de textos através da utilização das redes neurais MLP e RBF.

As redes MLPs são redes supervisionadas e consiste de unidades sensoriais que constituem a camada de entrada, uma ou mais camadas escondidas de nós computacionais e uma camada de saída de nós computacionais. O sinal de entrada se propaga para frente através da rede camada por camada. Apresenta um processo de aprendizagem supervisionado, no qual os padrões de treinamento são apresentados à

rede e com base nos erros obtidos são realizados ajustes nos pesos sinápticos, cuja finalidade é diminuir os erros nas próximas iterações. O principal algoritmo de treinamento das redes MLPs é o algoritmo de retropropagação (*back-propagation*) proposto no início por (Werbos, 1974) e reinventado diversas vezes, até se tornar popular com *Rumelhart* em 1986, (Rumelhart, 1986).

As redes RBF são constituídas de uma camada de entrada, apenas uma camada oculta (diferentemente das MLPs que podem ter uma ou mais camadas ocultas) que é responsável pela transformação não linear do espaço de entrada para o espaço oculto, sendo este último na maioria das vezes um espaço de alta dimensionalidade. E por fim a camada de saída que é linear e fornece uma resposta ao estímulo gerado pela aplicação dos dados de entrada, pela camada de entrada. O aprendizado de uma rede RBF é equivalente a encontrar uma superfície em um espaço multidimensional que melhor se ajuste ao conjunto de dados de treinamento, sendo o critério para “melhor ajuste” medido por algum critério estatístico.

Apesar dos modelos de redes MLP e RBF serem utilizados praticamente nas mesmas aplicações (aproximação de funções não lineares e classificação de padrões), existem algumas diferenças importantes entre os dois modelos. Fernando (1995) em sua tese de mestrado fez um comparativo bem detalhado sobre as principais diferenças entre estes modelos e em (Haykin, 2000) é possível encontrar mais detalhes sobre as redes MLP e RBF.

Uma das vantagens na utilização das redes neurais em aplicações envolvendo a classificação de textos é a redundância na representação de informações, o que torna o sistema tolerante às falhas. Uma desvantagem é a dificuldade de se trabalhar com um grande número de variáveis, o que pode tornar os algoritmos de treinamento computacionalmente exaustivos. De fato (Wiener *et al.*, 1995) afirma em seu trabalho que é possível obter bons resultados em aplicações envolvendo classificação de textos e redes neurais, entretanto uma redução no número de características do modelo se faz necessário.

4.1.2.4 Métricas de Avaliação

Diversas métricas podem ser utilizadas para realizar a avaliação de um classificador. (Witten *et al.*, 2011) definem a acurácia e a taxa de erro como medidas mais comuns para avaliar um modelo de classificação. Já (Sebastini, 1999) define que a

eficácia de um classificador pode ser avaliada em termos de precisão e *recall*. Estas medidas são estimativas dos percentuais de acertos e erros do modelo de predição da classe na apresentação de novos exemplos e as mesmas podem ser calculadas a partir de uma estrutura conhecida como matriz de confusão. A matriz confusão é apresentada na forma de tabela e para um problema envolvendo duas classes denominadas classe positiva e classe negativa, a mesma pode ser apresentada da seguinte forma:

| Classes | | Prevista | |
|---------|----------|----------------------------|----------------------------|
| | | Positiva | Negativa |
| Real | Positiva | TP (Verdadeiros Positivos) | FN (Falsos negativos) |
| | Negativa | FP (Falsos Positivos) | TN (Verdadeiros Negativos) |

Tabela 4.1 - Matriz Confusão

A matriz confusão apresentada acima indica quatro variáveis, ou seja, quatro possibilidades de acertos e de erros do classificador: Verdadeiros Positivos (TP) representa o número de registros da classe positiva preditos corretamente pelo modelo, Verdadeiros Negativos (TN) representa o número de registros da classe negativa preditos corretamente pelo modelo, Falsos Positivos (FP) representa o número de registros da classe negativa preditos como sendo da classe positiva pelo modelo e os Falsos Negativos (FN) representa o número de registros da classe positiva preditos pelo modelo como sendo da classe negativa. Para o caso de aplicações envolvendo mais de duas classes, considera-se cada classe como positiva e as demais como negativas.

A partir dos valores da matriz confusão diversas medidas de erro podem ser derivadas: acurácia, precisão, *recall* e a medida F (*F-Measure*) (Goadrich *et al.*, 2006).

A acurácia é a proporção dos documentos classificados corretamente (positivos e negativos) sobre o número total de documentos. De uma maneira geral quanto melhor o classificador, mais alta deve ser a acurácia. Entretanto deve-se ter cuidado ao utilizar esta métrica quando o *corpus* não possui a mesma quantidade de elementos por classe, ou seja, quando uma classe ocorre significativamente mais do que a outra, geralmente o classificador tende a obter uma alta acurácia classificando todos os exemplos na classe majoritária. Quando isto ocorre é importante utilizar outras medidas para avaliar o sistema (Goadrich *et al.*, 2006).

$$Acuracia = \frac{TP + TN}{TP + TN + FP + FN} \quad (4.4)$$

Precisão é a proporção dos documentos classificados corretamente como positivos. Pode-se dizer então que uma alta precisão significa que a maioria dos documentos classificados como positivos foram identificados corretamente. *Recall* mede a capacidade de reconhecer os documentos positivos, ou seja, pode-se dizer então que a maioria dos documentos positivos foram encontrados. A precisão é o número de documentos classificados como positivo divididos pelo número total dos documentos classificados como positivos, já *recall* é o número documentos corretamente classificados como positivo divididos pelo número total de documentos que realmente são positivos. As fórmulas 4.5 e 4.6 representam estas medidas.

$$Precisao = \frac{TP}{TP + FP} \quad (4.5)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.6)$$

Fragos, *et al.*, (2005) afirmam que as medidas precisão e *recall* podem ser enganosas quando examinadas separadamente. Uma precisão elevada geralmente significa sacrificar o *recall* e vice e versa. Os autores afirmam ainda que quando as medidas precisão e *recall* são sintonizadas para obter o mesmo valor, então este valor é chamado *break-even* (ponto de equilíbrio) do sistema e este *break-even* vem sendo muito utilizado em avaliações de sistemas de classificação de textos.

A *F-measure* (ou medida F) é a média harmônica entre a precisão e o *recall* e é definida pela equação 4.7.

$$F = \frac{2 * Precisao * Recall}{Precisao + Recall} \quad (4.7)$$

Precisão e *recall* capturam dois importantes aspectos do desempenho do classificador e a medida F combina estes dois aspectos em uma única medida, atribuindo importância igual a ambas as medidas. Quanto mais alto o valor da medida F melhor deve ser a performance do sistema.

5. Descrição do Sistema de Previsão do Mercado

Conforme citado anteriormente a previsão do mercado financeiro ainda é uma tarefa difícil devido à presença de dados ruidosos, não estruturados e com alto grau de incertezas. Diferentes modelos foram propostos nos últimos anos com diferentes tipos de dados (numéricos e textuais) para tentar entender o comportamento do mesmo. Não existe um consenso sobre quais são as melhores técnicas ou quais são os melhores dados a serem utilizados, entretanto é conhecido que muitos investidores fazem uso de notícias (principalmente macroeconômicas) em suas tomadas de decisão. Sendo assim a abordagem utilizada para a finalidade deste trabalho (previsão diária da tendência do principal índice do mercado de ações brasileiro) é a classificação de notícias financeiras através de técnicas de mineração de textos.

Vários estudos envolvendo este tema foram revisados no capítulo 2 deste trabalho, tendo também a apresentação de uma arquitetura geral dos modelos revisados. O modelo proposto segue a arquitetura geral apresentada no capítulo 2, porém com algumas modificações, conforme visualizado na figura 5.1 abaixo.

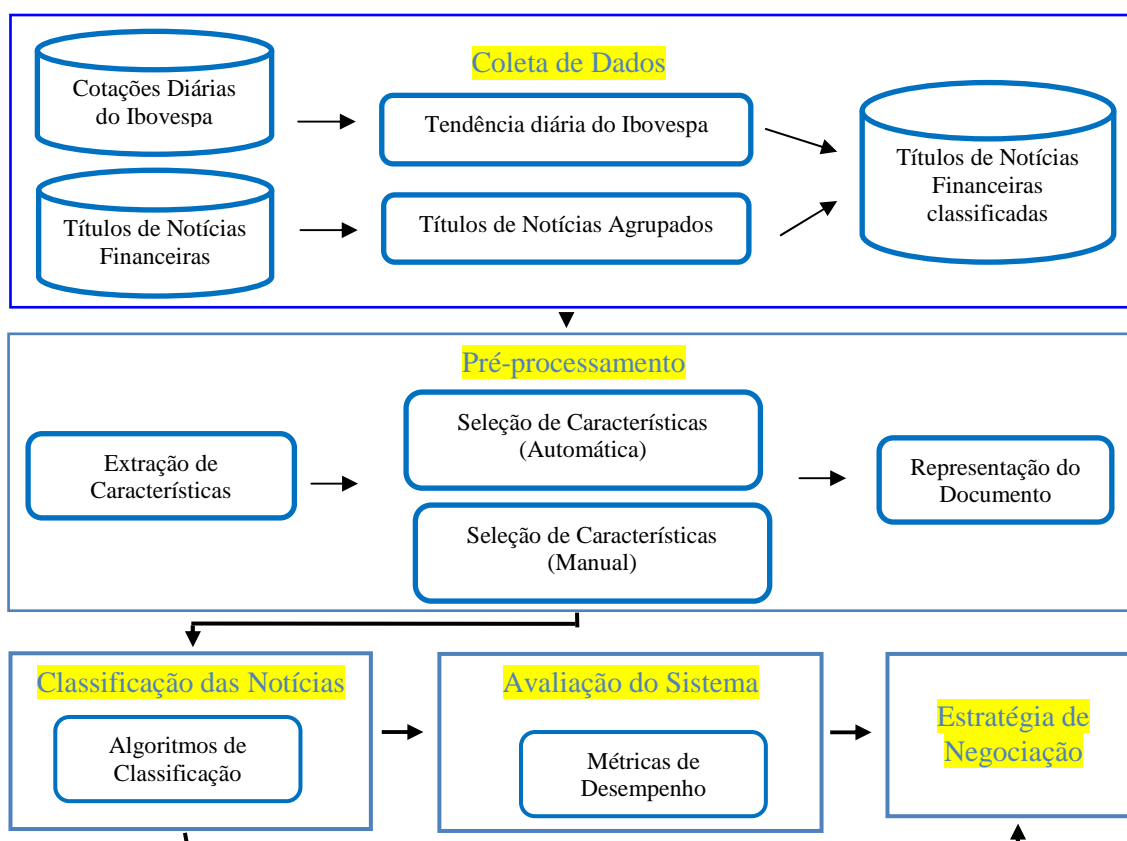


Figura 5.1 Etapas da Metodologia do sistema proposto

De uma maneira geral a metodologia proposta acima é constituída de cinco etapas: coleta de dados, pré-processamento, classificação das notícias, avaliação do sistema e por fim estratégia de negociação.

Para a implantação de todas as etapas da metodologia proposta foram realizados dois experimentos. No primeiro experimento (denominado nesta dissertação de Experimento1 ou Exp1) foi utilizado um software comercial chamado *Poly Analyst*¹³ (PA) desenvolvido pela empresa *Megaputer Intelligence Inc.* É um sistema de mineração de dados e mineração de texto que possibilita ao usuário executar diferentes tarefas desde as mais simples, como a importação, limpeza e manipulação de dados, a outras mais complexas como a modelagem de dados, avaliação, visualização de resultados e geração de relatórios. O software é composto de vários programas que dão suporte às tarefas citadas. E estes programas por sua vez possuem diferentes nós que fazem o tratamento dos dados. Sendo assim as tarefas sugeridas na metodologia proposta (com exceção das etapas de coleta de dados, avaliação do sistema e estratégia de negociação) foram executadas em um programa denominado *Analytical Client* (programa utilizado para a criação e execução de projetos de análises de dados) e no nó *Text Analysis* (nó constituído de vários módulos que dão suporte as tarefas de análise de dados textuais).

No segundo experimento (denominado Experimento2 ou Exp2) foram utilizadas duas linguagens de programação (*Mathematica* e *Matlab*) para implementar todas as etapas do modelo proposto, ou seja, um programa foi desenvolvido utilizando a linguagem de programação *Mathematica* para implementar as duas primeiras etapas do modelo proposto (coleta de dados e pré-processamento das notícias), enquanto as demais etapas (construção do classificador, avaliação do sistema e estratégia de negociação) foram implementadas no *Matlab*. Cabe ressaltar que a escolha por estas ferramentas se deu pela razão de que ambas as ferramentas são linguagens de programação que suportam a criação de novas funções e procedimentos, abrindo espaço para a completa edição dos mesmos, permitindo assim a implementação de todas as etapas necessárias para o desenvolvimento deste trabalho. Outra vantagem a ser citada é o fato das mesmas apresentarem características bem específicas para os propósitos escolhidos. No caso do *Mathematica* podemos afirmar que o mesmo é um software eficiente para trabalhar com a manipulação de textos e páginas da *web*. Enquanto o

¹³ *Poly Analyst Data Analysis*, Megaputer Intelligence, Inc. (<http://www.megaputer.com>).

Matlab apresenta várias *toolboxes* onde diversos recursos para a utilização de vários algoritmos de classificação se fazem presentes.

É importante destacar que a diferença entre os dois experimentos citados se encontra apenas nas etapas de pré-processamento e classificação das notícias. A ideia foi justamente avaliar a capacidade preditiva da tendência do índice do mercado de ações brasileiro através de duas metodologias de pré-processamento de textos diferentes e de algoritmos de classificação distintos.

Detalhes destes experimentos assim como uma explicação de todas as etapas da metodologia proposta são apresentados nas próximas seções.

5.1 Coleta de Dados

Dois tipos de dados são necessários para a realização do estudo proposto: (1) séries temporais contendo as cotações do fechamento diário do Ibovespa, (2) notícias financeiras contendo data e hora da publicação.

5.1.1 Base de dados de Séries Temporais Financeiras

As séries temporais contendo as cotações diárias do fechamento do Ibovespa foram adquiridas através de uma conceituada empresa de serviço de informações para o mercado financeiro. A série consiste não só de cotações diárias com valores de fechamento, mas também de valores de abertura, máximo, mínimo (para detalhes ver capítulo 3) e volume negociado (que corresponde ao volume financeiro de transações feitas em um determinado dia) no período de fevereiro de 2010 a junho de 2011. Vale ressaltar que neste estudo em questão apenas o valor diário do fechamento do Ibovespa foi utilizado. Um gráfico contendo os valores utilizados é apresentado na figura 5.2.

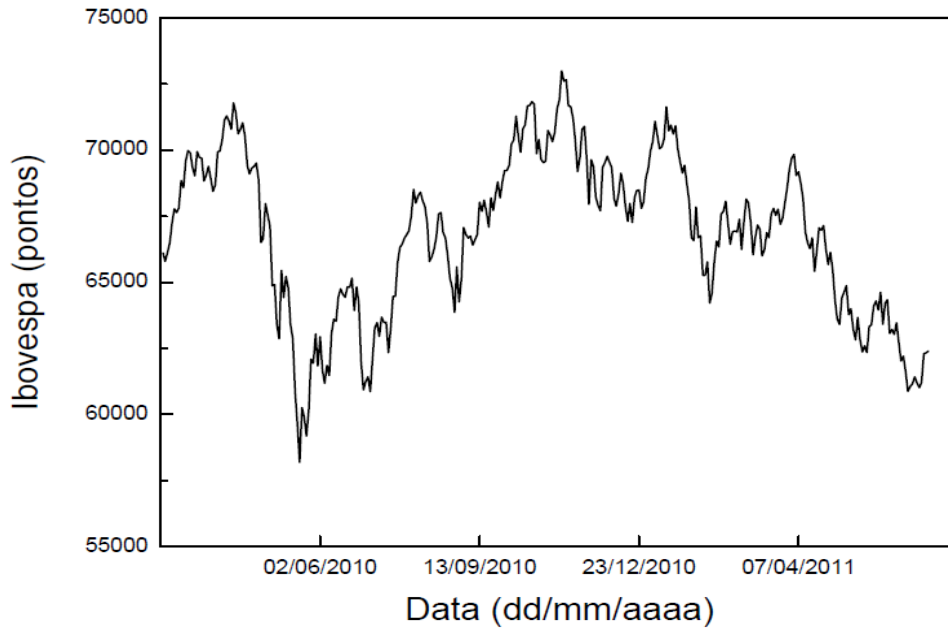


Figura 5.2 Série diária do fechamento do Ibovespa (fevereiro de 2010 a junho de 2011)

Conforme citado anteriormente (seção 2.1) do ponto de vista dos investidores de Bolsa de Valores as tendências das séries temporais financeiras são muito mais informativas do que o valor das ações, ou mesmo do índice Bovespa. A extração de tendência em uma série temporal pode ser feita de maneira simples considerando apenas o valor de fechamento consecutivo das ações ou índices, ou mesmo através de algoritmos mais sofisticados, como a segmentação linear *piecewise* (Keog, *et al.*, 2001). Neste trabalho em específico optou-se por extrair as tendências diárias do índice considerando o cálculo da variação relativa do Ibovespa, conforme a fórmula 5.1.

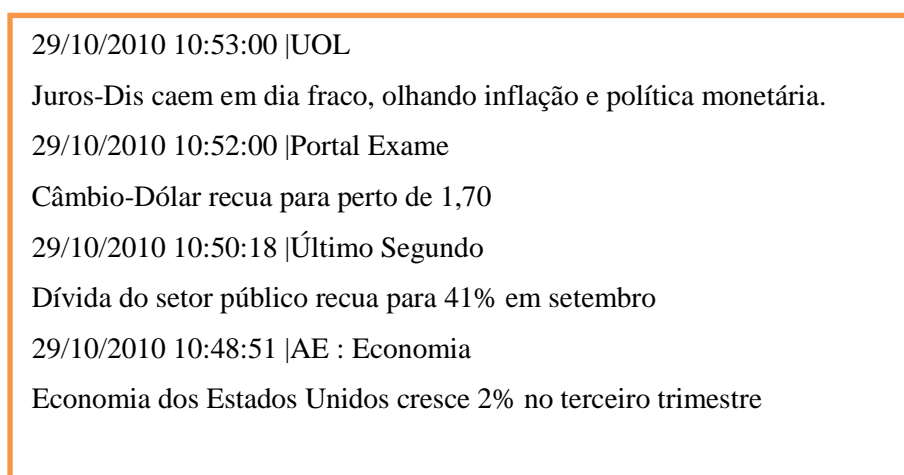
$$tendencia(t) = \frac{fechamento(t) - fechamento(t-1)}{fechamento(t-1)} * 100 \quad (5.1)$$

Onde t é um dia de negociação do Ibovespa e *fechamento* está relacionado ao último valor assumido pelo Ibovespa no horário regular do pregão.

5.1.2 Base de dados de Notícias Financeiras

Os outros dados necessários para o estudo, ou seja, notícias financeiras contendo data e hora de publicação foram adquiridas através dos principais sites de notícias econômicas (*Infomoney, Ig, Exame, Estadão, G1*, entre outros) do Brasil. Dentro da classificação dos tipos de notícias proposta por (Leinweber, 2009), as notícias utilizadas foram classificadas na categoria 1 (para detalhes ver capítulo 4), ou seja são notícias produzidas por grandes mídias e por fontes confiáveis e geralmente transmitidas através de jornais, rádios e televisão.

Para a tarefa de baixar estas notícias da internet um programa foi implementado no *Mathematica*. Este programa faz o *download* destas notícias e armazenam em um arquivo *txt* apenas os títulos¹⁴, hora e data de publicação e a fonte (ou site) da notícia. As demais informações contidas na página *web* são descartadas (como por exemplo, código *html*, figuras, entre outros). A figura 5.3 apresenta as informações armazenadas pelo programa.



```
29/10/2010 10:53:00 |UOL
Juros-Dis caem em dia fraco, olhando inflação e política monetária.
29/10/2010 10:52:00 |Portal Exame
Câmbio-Dólar recua para perto de 1,70
29/10/2010 10:50:18 |Último Segundo
Dívida do setor público recua para 41% em setembro
29/10/2010 10:48:51 |AE : Economia
Economia dos Estados Unidos cresce 2% no terceiro trimestre
```

Figura 5.3 Exemplo de Títulos de Notícias

Conforme pode ser observado, a figura acima contém as seguintes informações: a data, horário, o site (ou jornal *on-line*) em que a notícia foi publicada e também o título da notícia contendo as informações financeiras (estas informações geralmente

¹⁴ Considerando que todo título está associado a uma notícia e que o mesmo representa um resumo da notícia, não faremos distinção entre títulos de notícias e notícias neste capítulo, entretanto cabe ressaltar que a base de dados utilizadas neste trabalho é composta apenas dos títulos das notícias.

estão relacionadas a indicadores econômicos, situação econômica do Brasil e outros países como os Estados Unidos, Europa e Ásia, informações sobre mercado de ações internacionais, informações sobre taxas de juros, taxas de câmbio, etc.). É possível visualizar que cada título de notícia está associado com a data e horário de sua publicação. Esta informação é fundamental na definição do conjunto de dados de textos (títulos financeiros) e suas respectivas classes conforme será explicado na próxima seção.

A escolha pela utilização somente dos títulos das notícias e não de toda a notícia publicada está fundamentada no fato de que o título da notícia contém um vocabulário bem definido contendo apenas a informação relevante. E com isto é possível ter um conjunto de dados textuais mais relevantes e com menos ruído.

Os títulos de notícias baixadas da internet compreenderam ao período de 23/02/2010 a 30/06/2011 totalizando 174993 títulos em 337 dias de negociação na BmfBovespa. Uma tabela com as estatísticas básicas diárias destes títulos pode ser visualizada abaixo.

| | |
|---|---------------|
| Total de Notícias (títulos) | 174993 |
| Média de Notícias (títulos) diárias | 458 |
| Máximo de Notícias (títulos) diárias | 635 |
| Mínimo de Notícias (títulos) diárias | 19 |

Tabela 5.1 - Estatísticas básicas dos títulos das notícias no período de 23/02/2010 a 30/06/2011

5.1.3 Atribuição de classes aos Títulos das Notícias

Na maioria dos estudos envolvendo classificação de textos o processo de atribuição de classe geralmente é feita por um especialista da área. Neste estudo este processo é feito automaticamente baseado na hora em que os títulos das notícias diárias foram publicados e no valor da tendência diária do Ibovespa. Baseado neste processo de atribuição automática das classes aos títulos das notícias, o sistema será capaz de responder as seguintes questões: (1) Quanto tempo o mercado leva para reagir à publicação de determinadas notícias? (2) É possível prever a tendência diária utilizando apenas notícias que foram publicadas até a abertura dos negócios na BmfBovespa ? (3) Ou o investidor faz uso das notícias até momentos antes do fechamento dos negócios?

Para muitos autores na literatura estas questões são muito importantes, porque determinar o tempo (*timing*) que o mercado leva para reagir à publicação de notícias financeiras é fundamental para obter boas previsões. Por exemplo, em (Lavrenko *et al.*, 2000) os autores contrariam a hipótese do mercado eficiente (teoria que diz que as notícias (ou informações) são incorporadas no preço das ações imediatamente após o momento em que são divulgadas) e afirmam que este *timing* é de cinco horas, ou seja, o mercado demora cinco horas para reagir à publicação de uma determinada notícia. Já (Gidóvalfi, 2001) afirma que este *timing* é muito menor, ou seja, é de apenas 20 minutos. Segundo o autor existe uma forte correlação entre as notícias e o comportamento do preço das ações durante os 20 minutos que antecedem à publicação da notícia e os 20 minutos seguintes.

Buscando responder às questões citadas acima, neste estudo em específico optou-se por definir duas estratégias para a definição das bases de dados textuais:

1 – Na primeira estratégia todas as notícias publicadas em um determinado dia de negociação da BmfBovespa e antes de sua abertura são utilizadas para prever a tendência de seu fechamento naquele dia. Por exemplo: notícias publicadas no dia 29/10/2010 de 00h00min até 10h00min foram utilizadas para prever a tendência do fechamento do Ibovespa do dia 29/10/2010.

2 – Na segunda estratégia foram utilizadas todas as notícias de um determinado dia publicadas até 1 hora antes do fechamento da BmfBovespa naquele dia. Por exemplo: notícias publicadas no dia 29/10/2010 de 00h00min até 16h00min foram utilizadas para prever a tendência do fechamento do Ibovespa do dia 29/10/2010.

Dentro deste contexto duas observações a serem feitas: A primeira está relacionada às notícias publicadas em feriados e finais de semana. Para estas notícias foram atribuídas o valor da tendência do próximo dia de negociação do Ibovespa, ou seja, se as notícias foram publicadas no dia 02 de um determinado mês e este dia não é um dia de negociação da BmfBovespa, então estas notícias afetarão o mercado do dia 03. A segunda observação está relacionada às notícias publicadas depois do horário de fechamento da BmfBovespa (ou seja, notícias publicadas no intervalo de 17:00 horas até as 23:59 horas). Estas notícias em uma grande maioria apenas repetem o que aconteceu ao longo do dia de negociação, portanto acredita-se não acrescentar nenhuma informação nova ou relevante. Sendo assim as mesmas foram excluídas do modelo proposto.

Na tabela 5.2 abaixo podem ser visualizadas as estatísticas básicas das notícias (títulos) utilizadas nas estratégias 1 e 2 citadas acima. É possível perceber que com as estratégias adotadas a quantidade de notícias (títulos) utilizadas reduziu significativamente quando comparadas à quantidade de notícias baixadas pelo programa (ver tabela 5.1).

| | Estratégia 1 | Estratégia2 |
|---|--------------|-------------|
| Total de Notícias (títulos) | 53833 | 125615 |
| Média de Notícias (títulos) diárias | 160 | 373 |
| Máximo de Notícias (títulos) diárias | 585 | 823 |
| Mínimo de Notícias (títulos) diárias | 29 | 72 |

Tabela 5.2 - Estatísticas básicas das notícias após estratégias adotadas

A partir das estratégias adotadas acima dois conjuntos de dados de títulos de notícias foram criados, um com os títulos agrupados conforme estratégia 1 (Base1 contendo 337 documentos¹⁵) e um segundo conjunto com os títulos agrupados conforme estratégia 2 (Base2 também contendo 337 documentos).

Outro ponto a ser discutido é a quantidade de classes a ser adotada para classificar os documentos (ou títulos de notícias). A maioria dos trabalhos envolvendo previsão da tendência do mercado acionário citados na literatura fazem uso de modelos de três classes (positiva, negativa e neutra), ((Lavrenko *et al.*, 2000), (Gidóvalfi, 2001)). Outros preferem adotar um modelo de apenas duas classes como em (Zhai *et al.*, 2007). Devido ao fato de não se ter na literatura uma referência de trabalho envolvendo a previsão do mercado acionário brasileiro e a metodologia de mineração de textos optou-se pela utilização dos dois modelos citados na literatura, ou seja, o modelo de duas e também de três classes.

Sendo assim os documentos relacionados às Base1 e Base2 foram primeiramente classificados em três classes (positiva, negativa e neutra).

Conforme citado anteriormente estas classes foram atribuídas de acordo com a tendência do Ibovespa no dia da publicação dos textos. Logo as seguintes regras foram aplicadas para determinação destas classes:

¹⁵ Um documento contém todos os títulos agrupados conforme as regras propostas nas estratégias adotadas e o mesmo está relacionado a um dia de negociação da BmfBovespa.

- ✓ Se tendência do dia de negociação analisado é maior que 0,5% significa então que os textos daquele determinado dia influenciaram positivamente o valor do índice, logo classe é definida como positiva.
- ✓ Se tendência do dia de negociação analisado é menor que 0,5% significa então que os textos daquele determinado dia influenciaram negativamente o valor do índice, logo classe é definida como negativa.
- ✓ Se a tendência oscilou entre 0,5% e -0,5% significa então que os textos daquele determinado dia não influenciaram o valor do índice, logo classe é definida como neutra.

No final do processo de classificação dos documentos para o modelo de três classes, a distribuição das classes para cada conjunto definido (Base 1 e Base 2) ficaram da seguinte forma: 101 (29,97%) documentos foram classificados como positivos, 126 (37,39%) como neutros e 110 (32,64%) foram classificados como negativos, totalizando 337 documentos que correspondem a mesma quantidade de dias de negociação do Ibovespa.

Cabe ressaltar que três limiares foram testados a fim de se encontrar um melhor limiar para a atribuição das classes. O valor 0,5% citado anteriormente foi escolhido porque permitiu a criação de um modelo com classes balanceadas conforme pode ser observado na tabela 5.3. Segundo (Mittermayer, 2004) este é um ponto importante, porque muitos classificadores têm dificuldades de trabalhar com um conjunto de treinamento onde as classes são desbalanceadas.

| | Dias Positivos | Dias Neutros | Dias Negativos |
|------|----------------|--------------|----------------|
| 0,3% | 136 | 68 | 133 |
| 0,5% | 101 | 126 | 110 |
| 1% | 65 | 207 | 65 |

Tabela 5.3 - Distribuição para modelo de três classes para três limiares diferentes

No caso do modelo de apenas duas classes as regras para atribuição destas classes foram as seguintes:

- ✓ Se tendência do dia de negociação analisado é maior ou igual a zero significa então que os textos daquele determinado dia influenciaram positivamente o valor do índice, logo classe é definida como positiva.

- ✓ Caso contrário, ou seja, se tendência do dia de negociação analisado é menor que zero significa então que os textos daquele determinado dia influenciaram negativamente o valor do índice, logo classe é definida como negativa.

Sendo assim a quantidade de documentos por classe ficou da seguinte forma: 179 (53,11%) documentos foram classificados como positivos e 158 (46,89%) documentos foram classificados como negativos.

5.2 Pré-processamento das Notícias

Conforme citado no capítulo 4 deste trabalho quando se está trabalhando com dados textuais é necessária à transformação dos mesmos em uma representação adequada a ser utilizada pelos algoritmos de aprendizagem. A etapa responsável por esta transformação é denominada pré-processamento e na metodologia proposta nesta dissertação esta etapa é constituída de três atividades principais¹⁶: extração de características, seleção de características e representação do documento.

Na atividade extração de característica a preocupação é extrair do texto apenas termos relevantes, ou seja, termos que de uma maneira geral podem descrever suficientemente os textos. Métodos como eliminação de *stop words*, algoritmos de *stemming* e outros se fazem necessários nesta atividade.

Após a extração de características a próxima atividade no pré-processamento é a seleção de características. Nesta etapa o principal objetivo é eliminar aqueles termos que tem pouca relevância nos textos e assim poder diminuir a dimensionalidade do espaço de características melhorando assim a eficiência do classificador.

A seleção de características pode ser feita automaticamente, neste caso a utilização de métodos estatísticos se fazem necessário para selecionar aqueles termos que são relevantes para o estudo. Ou a mesma pode ser feita manualmente, neste caso um especialista da área define quais são os termos que são relevantes para o estudo.

A representação do documento é a última atividade a ser realizada dentro do pré-processamento. A mais comumente utilizada é a representação *bag of words* (onde cada documento é representado por um vetor de valores numéricos e cada valor indica a

¹⁶ Todas as três atividades foram explicadas no capítulo 4 desta dissertação.

importância do seu respectivo termo no documento). O valor ou peso destes termos pode ser calculado através de métodos estatísticos.

Como neste trabalho foram realizados dois experimentos diferentes, Exp1 e Exp2, optou-se por descrever a etapa de pré-processamento para ambos os experimentos em seções separadas, sendo assim os passos do pré-processamento para o Exp1 é apresentado na seção 5.2.1, enquanto os passos relacionados ao pré-processamento realizado no Exp2 é apresentado na seção 5.2.2.

5.2.1 Pré-processamento no Exp1

Neste experimento o pré-processamento foi realizado dentro do contexto do módulo denominado *Linear Classification* que faz parte do nó *Text Analysis* (conforme citado na introdução deste capítulo) disponível na ferramenta utilizada (PA). Por se tratar de uma ferramenta comercial, o *Poly Analyst* não explicita as técnicas envolvidas neste módulo, entretanto alguns conceitos estão disponíveis na documentação da ferramenta.

O módulo *Linear Classification* desenvolve um modelo de classificação dependente de um atributo estruturado, através da utilização de uma coluna independente dos textos. Para a etapa de pré-processamento o módulo disponibiliza dois parâmetros: *Use Stop List* e *Binary Word Count*.

O parâmetro *Use Stop List* pode ser utilizado ou não. Uma vez utilizado todas as palavras contidas em uma *stop list* são ignoradas durante a execução do módulo. Neste trabalho optou-se pela utilização do mesmo. Sendo assim uma lista contendo 1570 *stop words* (preposições, pronomes, artigos e outras classes de palavras auxiliares que não contribuem para o processo de classificação conforme citado no capítulo 4) em português foi adicionada à ferramenta, uma vez que a *stop list* disponível é específica para a língua inglesa.

O parâmetro *Binary Word Count* também pode ser utilizado ou não. Caso seja selecionado, a ferramenta armazena as palavras chaves de forma booleana em uma tabela, ou seja, se a palavra existe no documento seu valor é um, caso contrário seu valor é zero. Caso ele não seja selecionado as palavras chave são representadas em uma tabela por frequência, ou seja, é armazenada na tabela a frequência de cada palavra chave de um documento. Optou-se pela utilização das duas representações (tabela

booleana e tabela por frequência). A ideia foi verificar se a forma de representação do documento pode de alguma forma influenciar os resultados do experimento.

Infelizmente a forma como é implementado a seleção de características neste módulo não é reportado na documentação da ferramenta, apenas é possível afirmar que estas características são selecionadas automaticamente através da frequência e da distribuição dos termos nos textos conforme citado em (Poly Analyst, 2007).

5.2.2 Pré-processamento no Exp2

Em relação ao Exp2 foi desenvolvido um programa usando a linguagem *Mathematica*, (conforme citado na introdução deste capítulo) para a etapa de pré-processamento. O programa desenvolvido primeiramente fez a leitura dos textos (no formato *txt*) a serem processados.

O primeiro passo do pré-processamento está relacionado à extração de características. Sendo assim, para esta atividade o programa implementado foi capaz de executar as seguintes tarefas:

- ✓ Eliminação da *stop words*: nesta tarefa foram eliminados os termos que correspondem a pronomes, conjunções, artigos e preposições.
- ✓ *Stemming*: nesta tarefa foram executadas duas modificações nos documentos: eliminação de formas verbais (substituição dos verbos para sua forma infinitiva: Por exemplo, “sobem”, “subiu” e “subiram” foram substituídos por “subir”) e remoção do plural (por exemplo: “altas” = “alta”, “estáveis” = “estável”, *etc.*).
- ✓ Criação de um pequeno dicionário: O dicionário criado consistiu apenas na substituição de uma pequena lista de termos de mesmo significado pelo seu sinônimo correspondente. Por exemplo, os verbos “alavancar”, “aumentar” e “avançar” foram substituídos pelo verbo “subir”. É importante ressaltar que a escolha dos sinônimos para o dicionário criado está dentro do contexto de mercado acionário, ou seja, em outro contexto estes termos poderiam não ter o mesmo significado.

Ao final desta atividade os documentos processados são armazenados no formato *txt* para processamento posterior.

O próximo passo do pré-processamento é a seleção de características. No programa implementado não é executado esta atividade, pois esta seleção é feita manualmente, ou seja, todas as palavras chave utilizadas neste experimento fazem parte

de uma lista de sentenças composta de duas a cinco palavras (total de 90 sentenças) fornecidas por um especialista que atua no mercado de ações brasileiro. Algumas destas palavras chave são: inflação alta, dólar comercial abre em alta, juros futuros em queda, mercado em desaceleração, bolsa em queda, etc. Dentro deste contexto é possível afirmar então que a única tarefa executada no programa desenvolvido foi à leitura do arquivo *txt* contendo estas palavras chave para processamento posterior.

O último passo do pré-processamento está relacionado à representação do documento. Para esta atividade foi implementado no programa desenvolvido a representação *bag of words*, que pode ser facilmente convertida em tabelas. Sendo assim documentos processados e palavras chave são convertidos em uma tabela atributo-valor, onde cada documento é um registro da tabela e cada palavra chave é um elemento do conjunto de atributos¹⁷ e os valores numéricos (ou pesos) destes elementos foram calculados através das seguintes medidas estatísticas (ambas implementadas no programa): frequência do termo (TF), frequência TF-IDF (ambas explicadas na seção 4.1.2.2) e por fim a frequência TF-CDF (Peramunetilleke *et al.*, 2001). A TF-CDF faz uso da frequência do termo (TF) e da frequência de discriminação da classe (CDF) de um determinado atributo. Sendo esta última calculada pela seguinte fórmula:

$$CDF_i = \frac{\max(CF_{i,classe 1}, CF_{i,classe 2}, CF_{i,classe 3})}{DF_i} \quad (5.2)$$

Onde CF_i é o número de textos que contém um determinado atributo em uma classe em particular e DF_i é o número de documentos contendo o atributo pelo menos uma vez. Sendo assim a soma das frequências de todas as classes para cada atributo é igual à quantidade de documentos que o atributo aparece. Logo, o peso final do atributo i , ou seja, a $TF-CDF_i$ é então calculada multiplicando-se a frequência do termo (TF_i) pela frequência de discriminação CDF_i .

O resultado final do programa responsável pelo pré-processamento são várias tabelas, onde cada uma apresentou uma configuração possível de pré-processamento, ou seja, tabelas resultantes de documentos processados como os métodos de eliminação de

¹⁷ Por se tratar de uma representação em tabelas, vamos considerar todas as palavras chave que fazem parte dos resultados do pré-processamento como atributos.

stop words, *stemming* e as três métricas de atribuição de pesos (TF, TF-IDF e TF-CDF) e tabelas resultantes de documentos processados com métodos de eliminação de *stop words*, *stemming* e criação de dicionário e as mesmas métricas citadas anteriormente (TF, TF-IDF e TF-CDF). É possível perceber que a única diferença entre as tabelas geradas é a utilização ou não do dicionário criado. Cabe ressaltar que esta diferença pode influenciar muito nos resultados, pois a utilização do dicionário implica em uma diminuição das palavras chave (atributos) e conseqüentemente em um aumento na frequência destes atributos.

Na tabela 5.4 pode ser visualizada uma tabela resultante do pré-processamento e uma possível configuração.

| | Atributo 1 | Atributo2 | ... | Atributo 90 | Classe |
|-------------|------------|-----------|-----|-------------|--------|
| D1 | 0.2 | 0.21 | | 0.24 | 1 |
| D2 | 0.3 | 0.11 | | 0.67 | 2 |
| ... | ... | ... | | ... | ... |
| D337 | 0.76 | 0.12 | | 0.54 | 1 |

Tabela 5.4 - Representação final dos documentos após a conversão tabela Atributo-Valor

Na tabela cima o número de registros é igual à quantidade de documentos utilizada (337 documentos), o número de atributos é igual à quantidade de palavras chave fornecida pelo especialista (90 sentenças), o peso destes atributos é o valor calculado pelas medidas estatísticas utilizadas e a classe de cada registro esta relacionada à classe atribuída a cada documento conforme as estratégias citadas na seção 5.1.3 (no caso da tabela citada, o modelo utilizado é o de apenas duas classes).

5.3 Classificação dos documentos

O relacionamento entre o conteúdo dos títulos das notícias e a tendência do valor de fechamento do Ibovespa é aprendido através de diferentes algoritmos de aprendizado de máquina.

Na abordagem de aprendizado de máquina os documentos já previamente classificados são divididos em dois subconjuntos distintos, definidos como treinamento e teste (ou validação). O conjunto de treinamento é utilizado para treinar o classificador

e encontrar padrões nos documentos, enquanto o conjunto de teste (não rotulado) é utilizado para avaliar o desempenho do modelo. Como a base de testes também é previamente rotulada é possível então medir a taxa de acerto do modelo, comparando-se o resultado obtido com a rotulação disponível na base de testes. Os subconjuntos são normalmente disjuntos para assegurar que as medidas obtidas utilizando o conjunto de teste, sejam de um conjunto diferente daquele utilizado para realizar o aprendizado tornando a medida estatisticamente válida.

Rezende (2003) afirma que vários métodos como *Holdout*, Amostragem Aleatória, *Leave One-Out*, *Bootstrap* e *K-Fold-Cross Validation* podem ser utilizados para assegurar que as medidas obtidas pelo conjunto de validação sejam estatisticamente válidas e avaliar assim o desempenho de um classificador. Neste trabalho o método utilizado foi o *K-Fold Cross Validation* (ou validação cruzada de k partições). Este método consiste em três passos: inicialmente a base de dados é dividida em k subconjuntos aproximadamente iguais; em seguida $K-1$ subconjuntos são utilizados para treinamento (estimar o modelo) e o subconjunto restante é utilizado para testar; estes dois procedimentos são repetidos K vezes (neste trabalho foi definido k igual a 10) utilizando sempre um subconjunto diferente para testar. A consequência disto é que todos os documentos estão disponíveis para teste. O resultado final é obtido pela média dos resultados em cada etapa.

Uma vez definida a técnica de validação do modelo de classificação, os conjuntos de treinamento e teste (subconjuntos gerados pela técnica de validação cruzada de K partições) são então utilizados como entradas para os algoritmos de classificação definidos nos experimentos adotados (Exp1 e Exp2). Cabe ressaltar que a técnica utilizada (validação cruzada de k partições) foi implementada no *Matlab*.

5.3.1 Classificação de documentos no Exp1

No Exp1 a etapa de classificação de documentos também é executada através do módulo *Linear Classification* (mesmo módulo responsável pelo pré-processamento no PA) e do módulo *Score node*.

O módulo *Linear Classification* treina um modelo para classificar automaticamente os textos através de dois algoritmos de aprendizado de máquina distintos: (1) baseado no algoritmo SVM - algoritmo que demora mais tempo a ser

executado, uma vez que o mesmo requer um processamento mais intensivo a níveis computacionais, entretanto apresenta melhores resultados; (2) baseado no algoritmo bayesiano Simples – algoritmo mais veloz, pois apresenta um tempo computacional menor, entretanto a acurácia nos resultados pode ser inferior ao SVM. (Poly Analyst, 2007).

Neste sentido optou-se pela criação de vários modelos variando o conjunto de dados (Base1 e Base2), os parâmetros do pré-processamento citados anteriormente (*Use Stop List* e *Binary Word Count*) e os dois algoritmos disponíveis na ferramenta (SVM e Bayesiano Simples).

Após o processamento do módulo, o resultado final é um modelo de classificação de textos que é utilizado como entrada para o outro módulo da ferramenta o *Score node*, que é responsável pela avaliação do modelo. Sendo assim modelo de classificação gerado anteriormente e conjunto de teste foram utilizados como entrada para a execução do módulo *Score node*.

Ao final da execução deste módulo, os resultados finais obtidos foram apresentados através de uma tabela com as seguintes características: primeira coluna apresentou o documento propriamente dito, na segunda coluna, a classe correta do respectivo documento e na terceira coluna a classe sugerida pelo algoritmo de classificação. A partir destes resultados foi gerada uma matriz confusão contendo as taxas de erro de cada classe. Matriz esta gerada no *Matlab* dado que a ferramenta não disponibiliza esta opção no módulo “*Score node*”.

5.3.2 Classificação de documentos no Exp2

Em relação ao Exp2 os algoritmos de classificação utilizados foram as Redes Neurais Artificiais MLP e RBF. Sendo assim vários modelos de redes neurais MLP e RBF foram implementadas através da *toolbox* de redes neurais do *Matlab*.

A utilização de uma rede neural artificial exige a escolha de uma série de parâmetros, que podem de alguma maneira influenciar diretamente os resultados da rede. Infelizmente não existe um método que determine de forma direta quais são os melhores parâmetros para cada tipo de aplicação. Na maioria das vezes a busca por estes parâmetros pode ser demorada.

Dentre os parâmetros de uma rede neural, um dos mais significativos é a quantidade de neurônios da camada escondida. Neste sentido neste trabalho optou-se

pela modelagem de diferentes redes variando este parâmetro de forma a verificar qual a influência do mesmo na taxa de acertos de cada rede neural modelada.

Dentro deste contexto os modelos de redes (MLP e RBF) executados no software *Matlab* foram estruturados em uma camada de entrada onde a quantidade de neurônios foi definida pela quantidade de características ou atributos definidos após o pré-processamento dos textos, uma camada escondida (cuja quantidade de neurônios variou entre 5 e 40, mais precisamente foram criadas redes com 5, 10, 25, 30 e 40 neurônios) e uma camada de saída com dois neurônios para o caso do modelo de duas classes e três neurônios para o caso do modelo de três classes. Estes neurônios representam a tendência diária do Ibovespa para um determinado dia de negociação.

Foram treinadas diferentes arquiteturas de redes MLP e RBF variando a quantidade de neurônios da camada escondida. O treinamento para ambas as arquiteturas foi do tipo supervisionado, onde os dados de entrada (dados textuais processados em etapas anteriores) e suas respectivas classes que fazem parte do conjunto de treinamento foram então apresentados para a rede formando um par, entrada-saída de treinamento.

Após o treinamento das redes, o conjunto de teste é utilizado para validar o modelo. Os resultados finais obtidos assim como no Exp1 foram armazenados em uma matriz contendo as taxas de acerto e erro para cada classe. Esta matriz foi utilizada no próximo módulo do programa desenvolvido responsável pela avaliação do sistema e apresentado na próxima seção.

5.4 Avaliação dos Sistemas

A eficiência da classificação dos modelos desenvolvidos foi avaliada através dos métodos discutidos na seção 4.3, ou seja, as seguintes métricas: acurácia, precisão, *recall* e medida F. Todas estas medidas foram implementadas no *Matlab* e seus cálculos foram efetuados a partir dos resultados obtidos na matriz confusão. Apesar do PA não ter implementado as métricas de avaliação citadas nos módulos utilizados, a ferramenta permite exportar os resultados em forma de tabela ou em arquivos *txt*, sendo assim foi possível avaliar os resultados gerados no PA com as métricas sugeridas e implementadas no *Matlab*. É importante ressaltar que todas estas medidas de avaliação foram estimativas a partir do método de validação cruzada utilizado.

5.5 Estratégia de Negociação

O sistema de negociação proposta nesta dissertação implementa uma estratégia direta e simples, através de um método de decisão de compra e venda de ações. É importante ressaltar que o objetivo principal desta dissertação está relacionado ao principal índice do mercado de ações brasileiro (Ibovespa), no entanto para a estratégia de negociação adotada aqui, optou-se pela negociação de empresas que mantêm uma alta correlação com o Ibovespa¹⁸. Estas ações muitas vezes correspondem às principais *blue-chips* da bolsa brasileira, como é o caso das ações das empresas Petrobrás (petr4), Vale do Rio Doce (vale5) e Gerdau (GGBR4), entre outras. O fato de existir esta forte correlação entre as empresas apontadas anteriormente e o Ibovespa significa que em muitas das vezes basta o Ibovespa subir (ou cair) que as ações destas empresas respondem similarmente. O gráfico apresentado na figura 5.4 mostra a similaridade entre o Ibovespa e o valor de fechamento das ações da Vale do Rio Doce (representada no gráfico como Vale5) no período de 23-02-2010 a 30-06-2011.



Figura 5.4 Gráfico comparativo

Cotações dos fechamentos diários do Ibovespa e Vale do Rio Doce

¹⁸ Vale ressaltar que o Ibovespa também pode ser negociado na BmfBovespa através da compra e venda de minicontratos do índice.

Para as recomendações de compras e vendas de ações foi levado em consideração à classificação final dos documentos obtida pelos classificadores no conjunto de validação utilizado. Valores como custos de negociação de ações na BmfBovespa não foram considerados no sistema de negociação proposto, uma vez que a definição destes custos é difícil de ser quantificada, pois os mesmos dependem de alguns fatores como corretora envolvida, entre outros custos.

O sistema de negociação proposto envolve algumas hipóteses, a saber:

1. Na abertura da BmfBovespa é possível comprar ações com o valor em torno daquele com a qual a ação fechou o último negócio realizado no dia anterior (ou valor de fechamento da ação).
2. Suponha que o sistema prevê que o índice vai subir no dia em questão então:
 - 2.1 Se o preço do *call de abertura*¹⁹ apresentar uma variação relativa com relação ao fechamento anterior menor que o limiar pré-definido, então compra-se na abertura e vende-se no fechamento.
 - 2.2 Se o preço do *call de abertura* apresentar uma variação relativa acima do limiar pré-definido então não é feito nada.
3. Suponha que o sistema prevê que o índice vai cair no dia em questão então:
 - 3.1 Se o preço do *call de abertura* apresentar uma variação relativa menor que o limiar pré-definido, então vende-se na abertura (sem ter ainda comprado) e compra-se no fechamento.
 - 3.2 Se o preço do *call de abertura* apresentar uma variação relativa acima do limiar pré-definido então nada é feito.
4. Suponha que o sistema prevê que nenhuma alteração vai ocorrer no índice, então nenhuma compra ou venda é realizada neste dia.

¹⁹ Call de abertura: período compreendido aos minutos que antecedem a abertura das negociações na BmfBovespa.

Em resumo, depois de cada dia todas as posições são fechadas, ou seja, nenhuma ação é armazenada em custódia. O resultado líquido ao final do dia é um lucro²⁰ ou prejuízo financeiro dependendo do caso específico. Outra vantagem desta estratégia é que a operação em questão é identificada como uma operação de *day-trade*²¹, nas quais as taxas de corretagem muitas vezes são menores.

Valores entre 0,1% e 1,0% foram utilizados como limiares para a realização das hipóteses propostas acima. A ideia foi tentar encontrar um valor ótimo que pudesse gerar um maior lucro possível.

²⁰ É importante ressaltar que lucro está relacionado ao valor obtido por uma operação de venda e compra de ações (ou seja, venda-compra) levando em consideração os custos de transação da operação realizada. Entretanto nesta dissertação em específico lucro está sendo considerado como uma rentabilidade sem considerar os custos de transação.

²¹ Day-trade: conjugação de operações iniciadas e encerradas em um mesmo dia de negociação com o mesmo ativo, cuja quantidade negociada tenha sido liquidada total ou parcialmente.

6. Resultados e Análises

Utilizando a metodologia proposta no capítulo 5, diferentes modelos de previsão foram desenvolvidos com a finalidade de prever o índice (Ibovespa) do mercado acionário brasileiro, principal objetivo deste trabalho. A partir da metodologia proposta dois experimentos (Exp1 e Exp2) foram à base para a criação destes modelos de previsão. Os resultados obtidos em ambos os experimentos foram analisados e o desempenho dos diferentes modelos de previsão desenvolvidos foi avaliado utilizando as medidas de avaliação citadas no capítulo 4. Uma estratégia de negociação também foi utilizada para avaliar a lucratividade dos resultados obtidos pela metodologia proposta nesta dissertação.

6.1 Base de Dados: Séries Temporais Financeiras e Notícias

Os dados utilizados neste trabalho foram os mesmos para ambos os experimentos, ou seja, séries temporais contendo as cotações do fechamento diário do Ibovespa e títulos de notícias financeiras baixadas dos principais provedores de notícias financeiras do Brasil relacionado ao período de fevereiro de 2010 a junho de 2011. Primeiramente as séries temporais foram processadas, ou seja, foi feito o cálculo das variações relativas do Ibovespa no período estudado cuja finalidade foi encontrar a tendência diária do Ibovespa. No próximo passo os valores das tendências encontrados foram discretizados conforme estratégias adotadas na seção 5.3.1. Um gráfico contendo as cotações do Ibovespa juntamente com a variação relativa do mesmo no período estudado pode ser visualizado na figura 6.1.

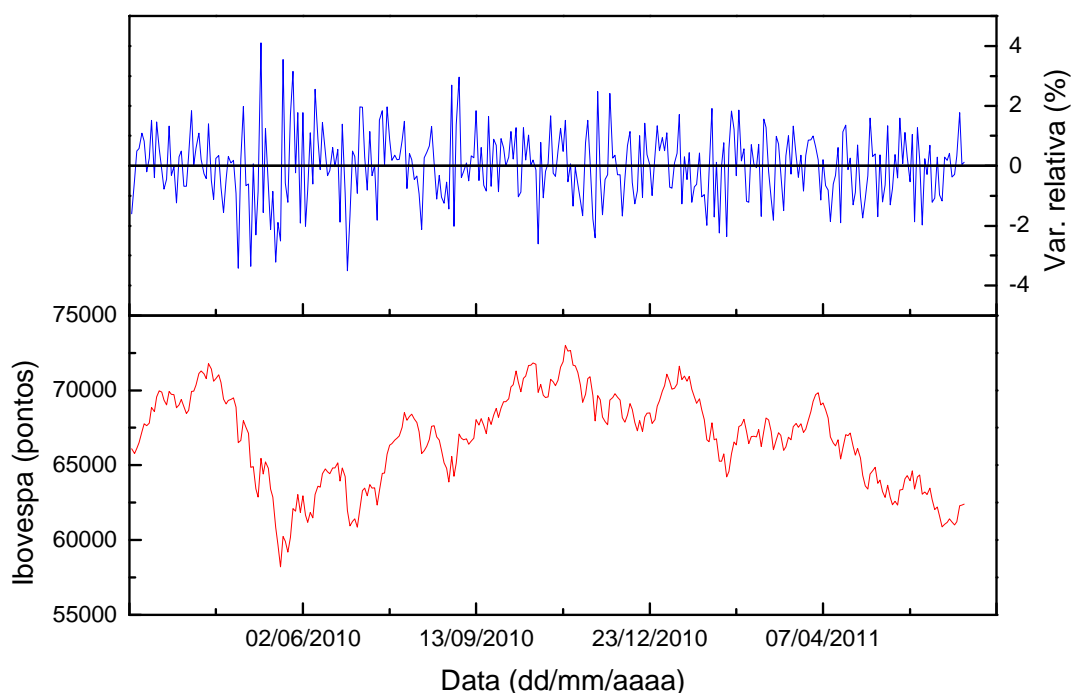


Figura 6.1 Cotações do fechamento do Ibovespa e Variação Relativa do Ibovespa

Através do gráfico acima é possível visualizar que a série temporal utilizada neste estudo (cotações dos valores de fechamento do Ibovespa) apresentou variações bruscas, sem uma tendência definida ao longo do período. Este ponto é muito importante, pois significa que os modelos desenvolvidos foram testados em condições complexas.

Em relação aos dados textuais utilizados neste trabalho, a base de dados criada para o período estudado consistiu de 174993 títulos de notícias conforme citado no capítulo 5. Estes títulos foram recuperados através dos principais sites de notícias econômicas do Brasil. Uma análise mais detalhada destas notícias demonstraram que existe uma forte sazonalidade na publicação das mesmas principalmente nos diferentes dias da semana. Um gráfico contendo as médias diárias destas publicações ao longo do período estudado pode ser visualizado na figura 6.2.

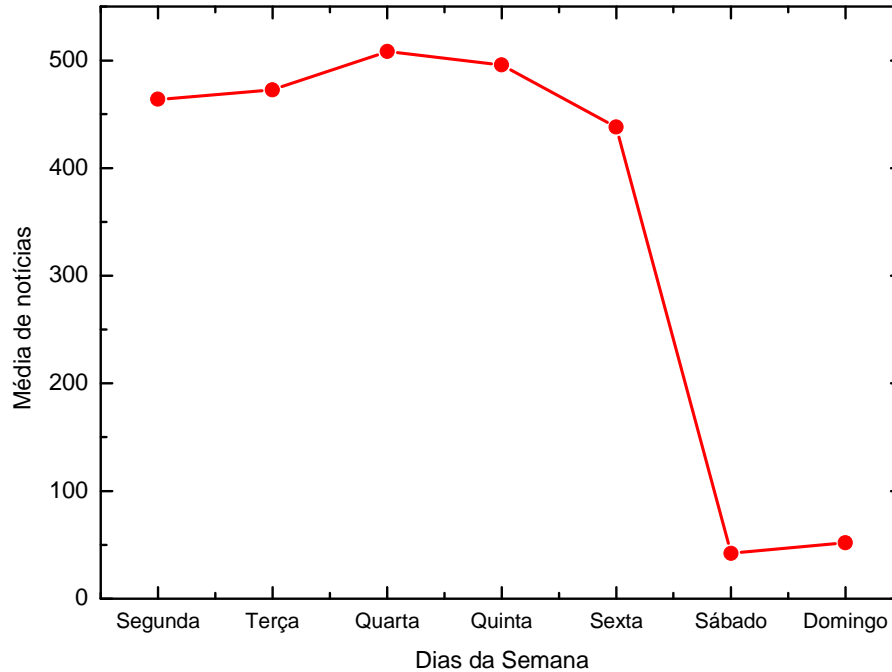


Figura 6.2 Gráfico da média de notícias diárias no período de fevereiro de 2010 a junho de 2011

Como era esperado, é possível visualizar no gráfico 6.2 que a quantidade de notícias publicadas nos dias úteis da semana (média entre 400 e 500 notícias diárias) é muito superior à quantidade de notícias publicadas nos finais de semana (média de 50 a 60 notícias). Outra observação a ser feita é que o pico de publicações de notícias ocorre na quarta e quinta feira e começa a cair a partir de sexta feira.

A partir deste conjunto de títulos de notícias (174993) foram definidos dois subconjuntos (Base1 e Base2) conforme estratégias adotadas na seção 5.1.3. Estas estratégias tiveram como base a hora de publicação da notícia. Cabe ressaltar que a hora do dia em que a notícia é publicada é um fato relevante para o entendimento da conexão entre o mercado de ações e as notícias, por isto a utilização da mesma para a definição das bases citadas.

As bases de dados definidas, Base1 e Base2 consistem de 53833 e 125615 títulos de notícias respectivamente. É possível perceber que o número de títulos de notícias para as duas bases geradas caiu drasticamente se comparadas à quantidade total de notícias recuperadas na internet (179499), principalmente a Base1, onde somente notícias publicadas até a abertura do pregão da BmfBovespa foram utilizadas.

6.2 Experimento 1. Resultados e Análise

Este experimento investigou a capacidade da metodologia de mineração de textos implementada na ferramenta PA em prever a tendência diária do índice (Ibovespa) do mercado acionário brasileiro. Os dados utilizados foram as Base1 e Base2, ambas citadas anteriormente e explicadas na seção 5.3.1 desta dissertação.

Para a realização deste experimento foram executados dois módulos: *Linear Classification*, responsável pelo pré-processamento dos textos e também pelo treinamento do classificador e *Score node* responsável pela validação do modelo gerado.

Três parâmetros foram testados e analisados na execução do módulo *linear classification*: *use stop list*, *binary count* e algoritmo de classificação (Naive e SVM). Logo, vários modelos foram desenvolvidos onde estes parâmetros foram modificados. Os diferentes modelos obtidos pelo *linear classification* alimentaram o módulo *score node* juntamente com o conjunto de validação (utilizado para avaliar o modelo gerado e definido em validação cruzada) e por fim os resultados finais da previsão foram obtidos. Uma análise sobre todos os resultados obtidos para os diferentes modelos gerados e validados são apresentados na próxima seção.

6.2.1 Resultados obtidos Experimento1

Seguindo os procedimentos descritos acima foram testados 16 modelos de previsão. Os resultados referentes a todos estes modelos podem ser visualizados nas figuras abaixo, sendo que a figura 6.1 está relacionada à Base1, enquanto a figura 6.2 está relacionada à Base2.

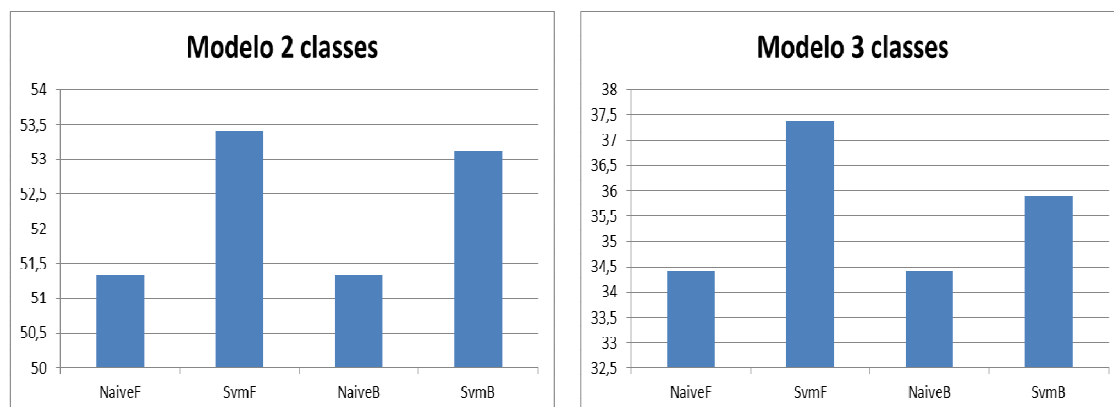


Figura 6.3 Acurácia dos classificadores para Base1

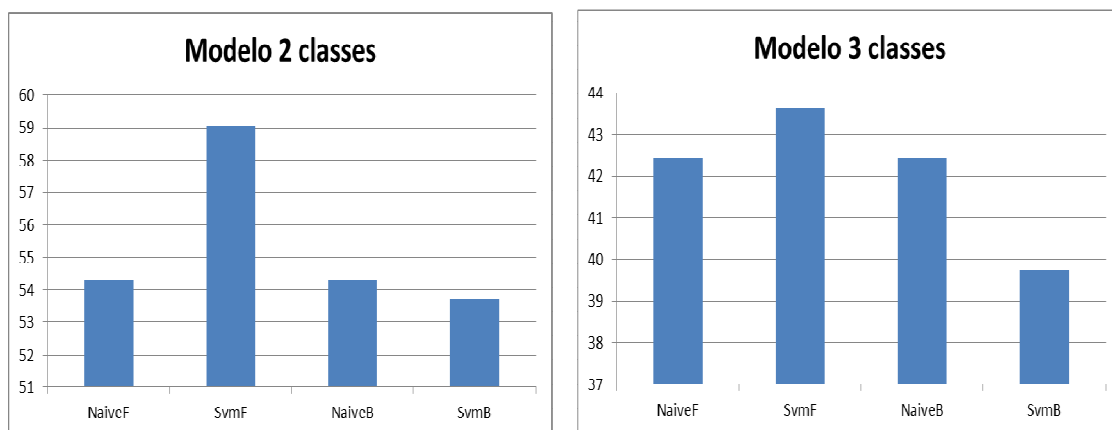


Figura 6.4 Acurácia dos classificadores para Base2

Nos gráficos das figuras acima, NaiveF e SVMF se referem a acurácia dos modelos gerados utilizando os algoritmo de classificação Naive e SVM respectivamente e a representação por frequência do documento para ambos os classificadores. Já NaiveB e SVMB estão relacionados à acurácia dos modelos gerados utilizando os mesmos algoritmos (Naive e SVM) porém com a representação booleana.

Analisando os gráficos anteriores é possível constatar alguns pontos relevantes que são discutidos a seguir. Todas estas conclusões parciais estão baseadas em correlações gerais entre os resultados obtidos (avaliada através da acurácia) e os diferentes parâmetros utilizados no Exp1.

- Os modelos de apenas duas classes apresentaram os melhores resultados tanto para a Base1 quanto para Base2. Esta conclusão pode parecer lógica dado que na classificação das bases de dados em apenas duas classes, aumentou-se o número de registro por classes e conseqüentemente os algoritmos de classificação conseguiram capturar melhor as estruturas inerentes nas amostras dos documentos pré-classificados.
- Em relação às duas bases de dados utilizadas (Base1 e Base2) foi possível constatar que os melhores resultados foram obtidos com a Base2. O fato desta base de dados apresentar um número maior de notícias (125647) pode ter contribuído para a obtenção destes resultados.
- Em relação à representação do documento, os melhores resultados foram obtidos com a representação por frequência. Este resultado já era esperado, uma vez que as bases de dados utilizados apresentaram uma alta frequência de palavras chave por documentos e a representação booleana é mais adequada apenas para textos

onde a frequência da maioria das palavras chave é extremamente baixa por documento, pois fornece pesos iguais a todas as palavras chave independente da frequência. (Poly Analyst, 2007).

- O algoritmo de classificação com maior acurácia para ambas as bases de dados (53,41% para a Base1 e 59,05% para a Base2) foi o SVM. Conforme citado em (Poly Analyst, 2007), o SVM realmente é um algoritmo mais preciso, o que justifica estes resultados, entretanto cabe ressaltar que apesar da simplicidade do *Naive*, os resultados obtidos pelo mesmo (51,34% para a Base1 e 54,30% para a Base2) não foram tão inferiores do que os resultados obtidos pelo SVM, principalmente se for levado em conta o tempo de execução do algoritmo onde o mesmo foi muito mais rápido do que o SVM.

Os resultados obtidos pelo classificador podem ser representados em uma estrutura conhecida como matriz confusão e a mesma pode ser utilizada para avaliar o desempenho do classificador conforme citado no capítulo 4. Sendo assim a matriz confusão relacionada ao modelo que obteve a melhor acurácia no Exp1 (Base2, modelo de duas classes, algoritmo SVM e representação por frequência) é apresentada na tabela abaixo.

| Classes | | Prevista | |
|---------|----------|----------|----------|
| | | Positiva | Negativa |
| Real | Positiva | TP = 120 | FN = 59 |
| | Negativa | FP = 79 | TN = 79 |

Tabela 6.1 - Resultado modelo de previsão para Base2

6.2.2 Análise e avaliação do Modelo no Exp1

Nesta seção foi avaliado o resultado apresentado na tabela 6.1 acima que está relacionado ao modelo de maior acurácia no Exp1. Os seguintes critérios de avaliação foram utilizados para esta finalidade: acurácia, precisão, *recall* e medida F.

De acordo com a matriz confusão (tabela 6.1) dentre os 337 documentos utilizados, 179 documentos pertencem à classe positiva e 158 pertencem à classe negativa. Dos 179 documentos classificados como positivo, o modelo previu 120 deles

corretamente como positivo e os demais 59 incorretamente classificados como negativo. Por outro lado dos 158 documentos classificados como negativo, 79 deles foram classificados corretamente como negativos e a mesma quantidade (79) classificados incorretamente como positivos. A partir destes valores foram calculados a acurácia total do modelo, recall para a classe positiva (*true positive rate*), *recall* para a classe negativa (*true negative rate*), precisão para classe positiva, precisão para a classe negativa e a medida F.

| | |
|----------------------------|---|
| Acurácia | $(120+79) / 337 = 59,05\%$ |
| Recall (Positivo) | $120/179 = 67,03\%$ |
| Recall (Negativo) | $79/158 = 50\%$ |
| Precisão (Positivo) | $120/199 = 60,30\%$ |
| Precisão (Negativo) | $79/148 = 53,37\%$ |
| Medida F (Positivo) | $(2*67,03*60,30)/67,03+60,30 = 63,48\%$ |
| Medida F (Negativo) | $(2*50*53,37)/53,37+50 = 51,63\%$ |

Tabela 6.2 - Medidas de Avaliação no Exp1

Através da acurácia do modelo de previsão é possível determinar que o modelo acertou corretamente 59,05% das tendências positivas e negativas. Analisando o *recall* para classe positiva (67,03%) e negativa (50%) é possível verificar que o modelo prevê melhor a tendência positiva do que a tendência negativa, uma diferença de 17,03%. O mesmo acontece com a precisão, ou seja, a precisão para tendência positiva é melhor do que a precisão para tendência negativa. Isto quer dizer que do número total de tendências positivas encontradas pelo modelo (120 de 199) a maioria delas realmente são positivas. Já entre o número total de tendências negativas encontradas pelo modelo apenas a metade delas foram realmente classificadas como negativas.

As medidas de desempenho (precisão e *recall*) podem ser enganosas quando examinadas separadamente. Uma precisão elevada geralmente significa sacrificar o *recall* e vice e versa. Sendo assim optou-se por utilizar também outro critério de avaliação que foi a medida F, que combina precisão e *recall*.

Conforme mencionado acima o *recall* para a classe positiva foi maior que o *recall* para a classe negativa, assim como a precisão que também apresentou um valor mais alto para a classe positiva. Ao examinar a medida F para a classe positiva

(63,48%) e para a classe negativa (51,63%) é possível afirmar com certeza que a capacidade do modelo em prever a classe positiva é muito melhor do que a classe negativa.

6.3 Experimento 2. Resultados e Análise

No Exp2 foram adotados novos procedimentos para alcançar os objetivos desta dissertação. Diferentemente do Exp1 quando uma ferramenta comercial foi utilizada, neste experimento foram desenvolvidos dois programas utilizando as linguagens de programação *Mathematica* e *Matlab* para a finalidade citada.

Neste experimento a associação entre as notícias e as tendências do Ibovespa foi construída a partir de dois modelos de Redes Neurais Artificiais, as redes MLP e redes RBF, ambas introduzidas no capítulo 4.

A utilização de uma rede neural artificial exige a escolha de uma série de parâmetros, que influenciam diretamente no resultado da rede. Para alcançar o objetivo desta dissertação foram implementadas no software *Matlab* vários modelos de redes MLP e RBF com diferentes configurações, buscando sempre um modelo com um resultado considerado como satisfatório. Conforme citado no capítulo 5 estes modelos foram estruturados da seguinte forma: uma camada de entrada onde a quantidade de neurônios foi definida pela quantidade de características ou atributos definidos após o pré-processamento dos textos, ou seja, modelos com 90 e 55 neurônios (é importante destacar aqui que o valor assumido por estes atributos foram definidos por três medidas estatísticas diferentes TF, TF-IDF e TF-CDF, sendo assim cada configuração foi testada com cada uma destas medidas); uma camada escondida onde a quantidade de neurônios assumiu os valores 5, 10, 25, 30 e 40 e uma camada de saída (representando a tendência diária do Ibovespa em um dia de negociação da BmfBovespa) com dois neurônios para os modelos de duas classes e três neurônios para os modelos de três classes.

6.3.1 Resultados obtidos Experimento2

Foram simuladas e testadas diferentes redes seguindo os procedimentos descritos anteriormente. De todas estas simulações com as diferentes configurações de RNAs utilizadas, aquelas que apresentaram o melhor desempenho juntamente com seus parâmetros (apenas os parâmetros que foram variados) são apresentadas na próxima

seção. É importante destacar que diferentemente do Exp1 onde todos os resultados gerados foram apresentados, para o Exp2 devido a grande quantidade de modelos gerados, apenas aqueles com melhor desempenho são apresentados nas figuras 6.5 (relacionada às redes MLP) e 6.6 (relacionada às redes RBF).

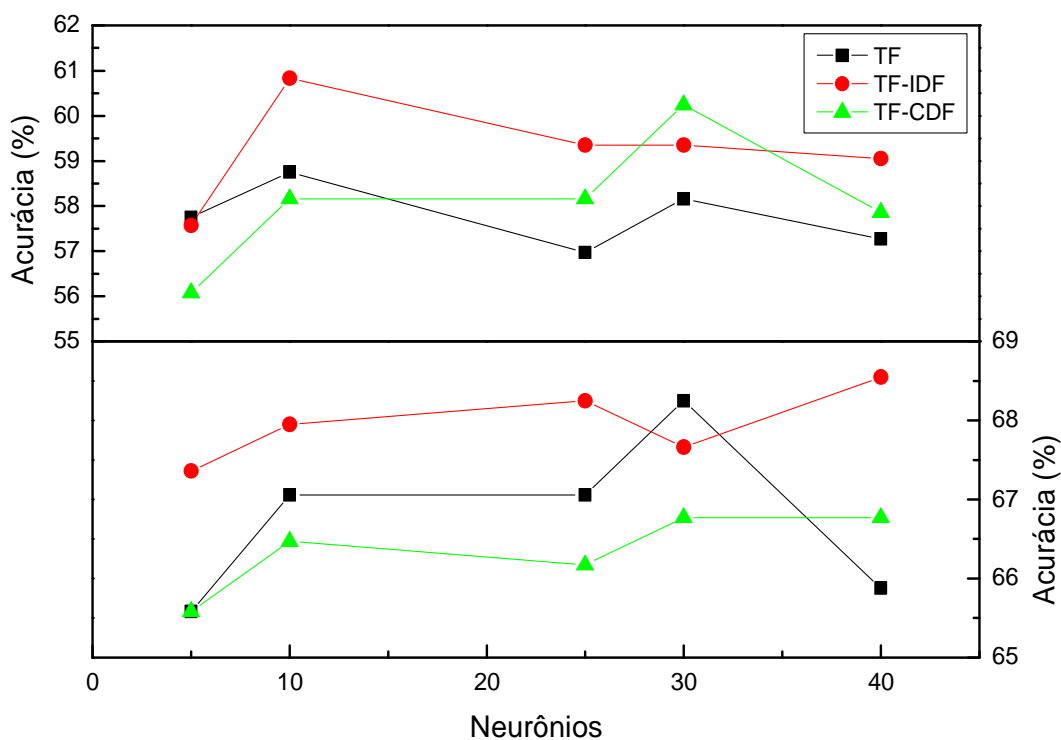


Figura 6.5 Resultados redes MLP para modelo de duas classes. Gráfico superior está relacionado à Base1, enquanto gráfico inferior está relacionado à Base2

A figura 6.5 apresenta dois gráficos plotados sobre o mesmo eixo de parâmetros. Ambos mostram a métrica de desempenho (acurácia) em função da quantidade de neurônios escondidos. As configurações dos modelos plotados nos gráficos foram a seguinte: gráfico superior (redes MLP; quantidade de neurônios variando entre 5 e 40; base de dados utilizadas foi a Base1; métricas estatísticas para atribuição de pesos foram a TF, TF-IDF e TF-CDF e por fim a quantidade de classes que foram duas) e o gráfico inferior a mesma configuração com exceção da base de dados que está relacionada à Base2.

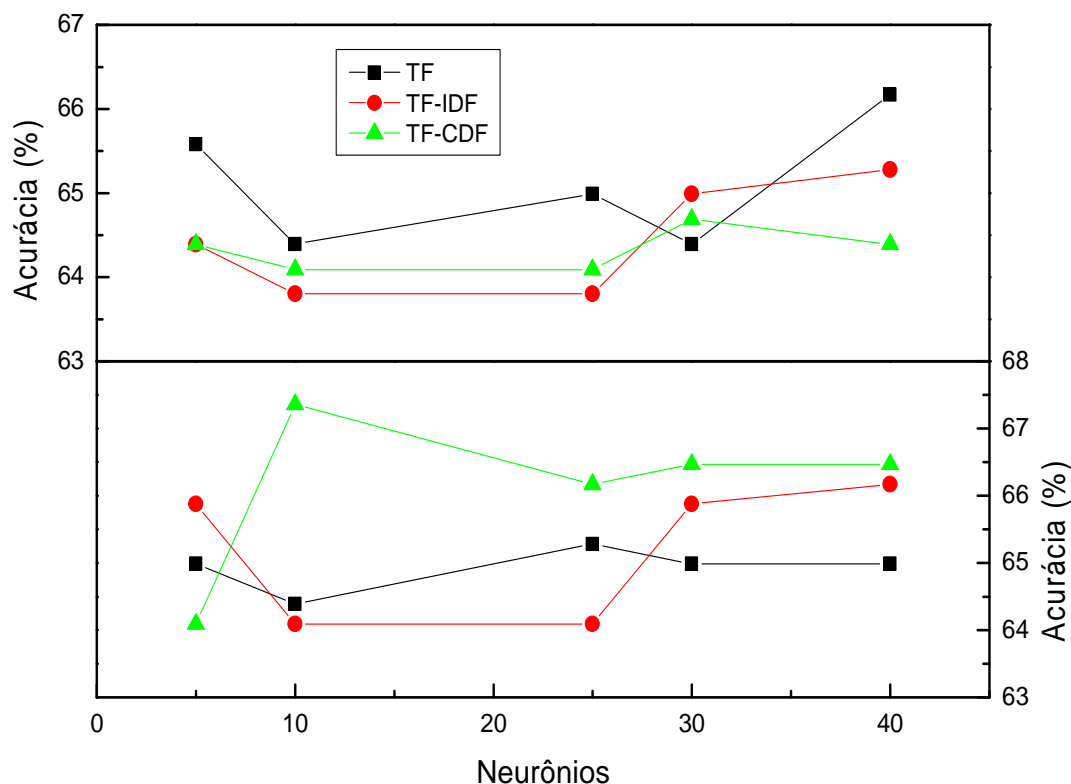


Figura 6.6 Resultados redes RBF para modelo de duas classes. O gráfico superior está relacionado à Base1, enquanto gráfico inferior se refere à Base2

No gráfico 6.6 têm-se as mesmas características citadas para o gráfico 6.5, porém com uma modificação no tipo de rede, ou seja, neste gráfico os resultados estão relacionados às redes RBF.

Analisando os resultados obtidos com todas as possíveis configurações de rede proposta neste trabalho e também os gráficos anteriores é possível constatar alguns pontos relevantes neste experimento e os mesmos são discutidos a seguir. Cabe ressaltar que todas estas conclusões parciais estão baseadas em correlações gerais entre os resultados obtidos (avaliada através da acurácia) e os diferentes parâmetros utilizados na modelagem das RNAs.

- O primeiro ponto a ser destacado é que analisando os gráficos acima não é possível verificar uma correlação entre os melhores resultados obtidos e a quantidade de neurônios, ou seja, o fato de aumentar a quantidade de neurônios da camada escondida não necessariamente significa em melhorar a acurácia do modelo.

- Quanto às medidas utilizadas para a atribuição de pesos às palavras chave ou atributos (TF, TF-IDF, TF-CDF) os resultados foram os seguintes: em relação ao modelo de rede MLP e ambas as bases de dados utilizadas (Base1 e Base2) os melhores resultados foram obtidos com a TF-IDF (Base1 = 60,83% com 10 neurônios na camada escondida e Base2 = 68,55% com 40 neurônios na camada escondida). Já em relação ao modelo de rede RBF e ambas as bases de dados utilizadas (Base1 e Base2) os melhores resultados foram obtidos com medidas diferentes para cada base de dados, ou seja, para RBF e Base1 foi a medida TF (66,17% com 40 neurônios) e para RBF e Base2 foi a medida TF-CDF (67,36% com 10 neurônios). Sendo assim não é possível definir exatamente qual medida de atribuição de pesos é melhor para todas as possíveis configurações analisadas, entretanto cabe ressaltar que assim como Sebastiani (1999), que mostrou que a TF pode ser muito eficiente, analisando os resultados obtidos aqui, podemos afirmar que apesar de sua simplicidade a TF apresentou bons resultados principalmente na comparação da mesma com a medida TF-IDF.
- Em relação aos modelos de redes MLP e RBF é possível afirmar que nenhum dos dois modelos conseguiu obter os melhores resultados para ambas as bases de dados utilizadas (Base1 e Base2), ou seja, em relação a Base1 o melhor resultado foi obtido com a rede RBF (66,17%), enquanto que para a Base2 o melhor resultado foi obtido com a rede MLP (68,55%).
- Em relação às bases de textos utilizadas (Base1 e Base2) foi possível verificar que a Base2 obteve resultados bem parecidos para ambas às arquiteturas de rede (MLP e RBF), ou seja, todos os resultados ficaram acima de 64%. Já em relação a Base1 somente os modelos RBF conseguiram resultados iguais ou superiores a 64%.
- Quanto à quantidade de classes utilizadas nos modelos gerados foi possível verificar que assim como no Exp1 os melhores resultados foram obtidos com duas classes. É importante destacar que os resultados obtidos com três classes variaram entre 40 e 50% para todas as possíveis configurações utilizadas neste trabalho. Por isto os mesmos não foram apresentados nesta seção.
- Em relação ao pré-processamento dos documentos, optou-se também pela utilização ou não de um dicionário conforme citado na seção 5.3.2. Pode-se afirmar que para modelos que fizeram utilização do dicionário no pré-

processamento os resultados obtidos foram inferiores quando comparados aos modelos que não fizeram uso deste dicionário. Acredita-se que este resultado ruim se deu devido ao fato de que mesmo que os termos utilizados tenham sido substituídos por seus sinônimos correspondentes e também dentro do contexto de mercado acionário, os mesmos podem ter um peso diferente, ou seja, tomando como exemplo as frases “Bolsa de Tóquio subiu” e “Bolsa de Tóquio disparou” podemos afirmar nas duas frases que a Bolsa de Tóquio subiu, entretanto o fato da Bolsa de Tóquio ter disparado pode ter mais relevância do que dizer simplesmente que a bolsa subiu. Sendo assim é possível concluir que mesmo que estas sentenças signifiquem a mesma informação, uma pode ter um peso maior que a outra.

Por fim, assim como no Exp1 o resultado obtido pelo modelo que apresentou a maior acurácia foi representado por uma matriz confusão. Vale ressaltar que a maior acurácia obtida foi com o modelo de redes MLP, Base2, 40 neurônios na camada escondida e medida TF-IDF (acurácia de 68,55%), entretanto optou-se por considerar como melhor resultado aquele obtido pelo modelo de rede RBF, Base1, 40 neurônios na camada escondida e medida TF (acurácia de 66,17%). A escolha por este modelo se deu devido a duas hipóteses:

1. Os melhores resultados obtidos tanto para Base1 (66,17%), quanto para Base2 (68,55%) foram bem semelhantes.
2. O fato da Base1 ser constituída de notícias agrupadas e divulgadas antes do horário de abertura da BmfBovespa pode ser considerado uma vantagem, ou seja, na abertura do pregão já se sabe qual será a tendência de fechamento do Ibovespa com uma acurácia de 66,17%. Considera-se então que esta informação é muito mais valiosa do que ter-se um modelo com uma acurácia de 68,55% apenas às 16 horas (lembrando que a Base2 é composta de notícias agrupadas até às 16 horas, ou seja, apenas há uma hora antes do fechamento da BmfBovespa).

Logo a matriz confusão apresentada a seguir está relacionado ao modelo de seguinte configuração: rede RBF, Base1, 40 neurônios na camada escondida e medida TF. Uma análise mais detalhada da matriz confusão citada é apresentada na próxima seção.

| Classes | | Prevista | |
|---------|----------|----------|----------|
| | | Positiva | Negativa |
| Real | Positiva | TP = 140 | FN = 39 |
| | Negativa | FP = 75 | TN = 83 |

Tabela 6.3 - Resultado modelo de previsão para Base1

6.3.2 Análise e avaliação do Modelo no Exp2

Nesta seção avalia-se o resultado apresentado na tabela 6.3 acima utilizando-se as mesmas medidas citadas na avaliação do Exp1, ou seja, acurácia, precisão, *recall* e medida F.

Como no Exp1 dentre os 337 documentos utilizados 179 deles pertencem à classe positiva e 158 a classe negativa. Dos 179 documentos pertencentes à classe positiva, o modelo previu 140 deles corretamente como positivo e os demais 39 incorretamente classificados como negativo. Dentre os 158 documentos pertencentes à classe negativa, 83 deles foram classificados corretamente como negativos e 75 classificados incorretamente como positivos. A partir destes valores foram calculados a acurácia total do modelo, *recall* para a classe positiva (*true positive rate*), *recall* para a classe negativa (*true negative rate*), precisão para classe positiva, precisão para a classe negativa e a medida F.

| | |
|----------------------------|---|
| Acurácia | $(140+83) / 337 = 66,17\%$ |
| Recall (Positivo) | $140/179 = 78,21\%$ |
| Recall (Negativo) | $83/158 = 52,53\%$ |
| Precisão (Positivo) | $140/215 = 65,11\%$ |
| Precisão (Negativo) | $83/122 = 68,03\%$ |
| Medida F (Positivo) | $(2*78,21*65,11)/65,11+78,21 = 71,06\%$ |
| Medida F (Negativo) | $(2*52,53*68,03)/68,03+52,53 = 59,28\%$ |

Tabela 6.4 - Medidas de Avaliação no Exp2

Através da acurácia do modelo de previsão é possível determinar que o modelo acertou corretamente 66,17% das tendências positivas e negativas. Analisando o *recall*

para classe positiva (78,21%) e negativa (52,53%) é possível verificar que o modelo prevê muito melhor a tendência positiva do que a tendência negativa, uma diferença de 25,68%. Por outro lado, analisando os valores da precisão é possível verificar que inversamente ao *recall*, a precisão para a classe negativa é melhor do que a precisão para a classe positiva.

Assim como no Exp1 a medida F (medidas da precisão e *recall* combinadas) também foi utilizada para avaliar o modelo. Como citado anteriormente o *recall* e a precisão apresentaram medidas inversas, sendo assim a decisão sobre qual categoria é melhor predita pelo modelo é uma decisão difícil. Entretanto analisando a medida F para a classe positiva (71,06%) e para a classe negativa (59,28%) é possível determinar com certeza que a capacidade do modelo em prever a classe positiva é muito melhor que a capacidade de prever a classe negativa.

6.4 Comparação entre os resultados

Nesta seção serão consideradas duas comparações diferentes, ou seja, primeiramente é apresentada uma comparação entre os resultados obtidos nos dois experimentos propostos nesta dissertação (Exp1 e Exp2). E em seguida outra comparação é apresentada, porém entre os resultados obtidos nesta dissertação (resultado de maior acurácia) e os resultados obtidos por (Faria *et al.*, 2009), onde o mesmo indicador foi previsto através de técnicas de redes neurais artificiais e alisamento exponencial adaptativo.

6.4.1 Comparação entre os resultados do Exp1 e Exp2

Conforme citado anteriormente para prever a tendência diária do Ibovespa neste trabalho foram executados dois experimentos denominados neste trabalho de Exp1 e Exp2. Dentro do contexto da metodologia proposta, a principal diferença entre estes dois experimentos foi a etapa de pré-processamento dos documentos e a etapa de classificação de notícias.

Analisando os resultados obtidos nos dois experimentos citados (tabela 6.5) e analisando as tarefas executadas nas duas etapas que diferenciaram um experimento do outro é possível destacar alguns pontos relevantes:

- O primeiro ponto a ser destacado é que os resultados obtidos no Exp1 (maior acurácia = 59,03%) foram inferiores aos resultados obtidos no Exp2 (maior acurácia considerada = 66,17%).
- Em relação ao pré-processamento dos documentos, o Exp1 tem uma vantagem em relação ao Exp2, pois como a ferramenta utilizada faz uso de medidas estatísticas para a seleção de características, a utilização de um especialista para fornecer as palavras chaves não se faz necessário. Entretanto acredita-se que a qualidade das palavras chaves obtidas pela ferramenta não representaram tão bem os documentos e que a inferioridade nos resultados pode ser consequência disto.
- Quanto aos algoritmos de classificação utilizados, não é possível fazer uma comparação direta entre os resultados obtidos pelos mesmos, dado que cada algoritmo tem a sua metodologia, entretanto é possível destacar dois pontos fundamentais. O primeiro é que o algoritmo SVM é composto de vários parâmetros que podem de alguma maneira influenciar os seus resultados como a constante de regularização, tipo de *kernel* (polinomial ou gaussiano), grau do polinômio (dependente da escolha do *kernel* polinomial), entre outros, portanto o módulo responsável pela classificação dos textos no Exp1 não permite a variação destes parâmetros, por isto nenhuma configuração diferente da proposta pela ferramenta (configuração *default*) foi utilizada. Já em relação às redes neurais utilizadas no Exp2, pode-se afirmar que as mesmas também fazem uso de uma série de parâmetros que de alguma maneira pode influenciar os resultados dos modelos gerados, sendo assim o fato de poder variar um dos seus principais parâmetros (quantidade de neurônios da camada escondida) pode ser considerado como um fator relevante na obtenção dos melhores resultados. Cabe destacar que recentemente os sistemas de otimização de modelos de redes neurais artificiais determinam automaticamente qual o modelo e método deve ser utilizado para cada aplicação específica (Trelea, 2003).
- E por fim é possível destacar que analisando a tabela 6.5 referentes às medidas precisão, *recall* e medida F é possível afirmar que os resultados obtidos pelo Exp2 superaram os resultados obtidos no Exp1 em todas as medidas utilizadas, o que confirma a superioridade da capacidade preditiva do Exp2.

| | Acurácia | Recall (Pos) | Recall (Neg) | Precisão (Pos) | Precisão (Neg) | Medida F (Pos) | Medida F (Neg) |
|----------|----------|-----------------|-----------------|-------------------|-------------------|-------------------|-------------------|
| Exp1 (%) | 59,05 | 67,03 | 50,00 | 60,30 | 53,37 | 63,48 | 51,63 |
| Exp2 (%) | 66,17 | 78,21 | 52,53 | 65,11 | 68,03 | 71,06 | 59,28 |

Tabela 6.5 – Comparação entre os resultados Exp1 e Exp2

Pos = Positivo e Neg = Negativo.

6.4.2 Comparação entre os resultados Exp 2 e (Faria et al., 2009)

Uma das propostas deste trabalho foi a comparação entre os resultados obtidos nesta dissertação e os resultados obtidos em (Faria *et al.*, 2009) onde o mesmo indicador foi previsto através de técnicas de mineração de dados.

A ideia principal foi verificar se modelos de previsão (neste caso em específico, previsão do Ibovespa) baseados em notícias financeiras e métodos de mineração de textos poderiam apresentar resultados superiores quando comparados a modelos de previsão baseados em mineração de dados (dados numéricos, muitas das vezes séries temporais de indicadores técnicos), como é o caso do modelo proposto por (Faria *et al.*, 2009) utilizados nesta comparação.

Analisando a acurácia dos dois modelos citados, ou seja, (Faria *et al.*, 2009) com acurácia de 60% e modelo proposto nesta dissertação com acurácia de 66,17% foi possível constatar que a utilização de informação textual (notícias macroeconômicas) em modelos de previsão do indicador da Bolsa de Valores de São Paulo melhora a performance dos modelos e conseqüentemente melhora a acurácia das previsões. Entretanto é importante ressaltar alguns pontos relevantes desta comparação:

- A previsão em ambos os trabalhos foram realizadas de forma diferentes, ou seja, no trabalho dos autores (Faria *et al.*, 2009) uma janela do tempo ($t-1$) com cotações diárias do fechamento do Ibovespa foram utilizadas para prever a tendência do fechamento do Ibovespa no dia t (dia seguinte). Nesta dissertação títulos de notícias diários foram utilizadas para prever a tendência diária do Ibovespa no mesmo dia em que os mesmos foram publicados.
- Os períodos analisados nos dois trabalhos foram diferentes (setembro de 1998 até abril de 2007) no trabalho dos autores (Faria *et al.*, 2009) e (fevereiro de 2010 a junho de 2011) nesta dissertação. Apesar desta diferença, pode-se afirmar que ambos os períodos apresentarem uma alta volatilidade.

Por fim conclui-se que levando em consideração as observações citadas anteriormente é possível afirmar que a utilização de notícias macroeconômicas como entradas para modelos de previsão do indicador da Bolsa de Valores de São Paulo pode apresentar resultados superiores, quando comparados a modelos onde somente dados numéricos são utilizados.

6.5 Resultados. Estratégia de Negociação

Conforme citado no capítulo anterior uma estratégia de compra e venda de ações da Vale do Rio Doce (Vale5) foi proposta nesta dissertação. A ideia foi verificar a lucratividade do modelo de maior performance obtido nos experimentos executados, ou seja, modelo realizado no Exp2 analisado na seção 6.3.2, cuja acurácia foi igual a 66,17%.

Algumas hipóteses foram levadas em consideração nas estratégias adotadas e diferentes limiares foram testados conforme citado na seção 5.5. Os resultados referentes a todos estes testes podem ser visualizados no gráfico abaixo, assim como uma análise dos mesmos.

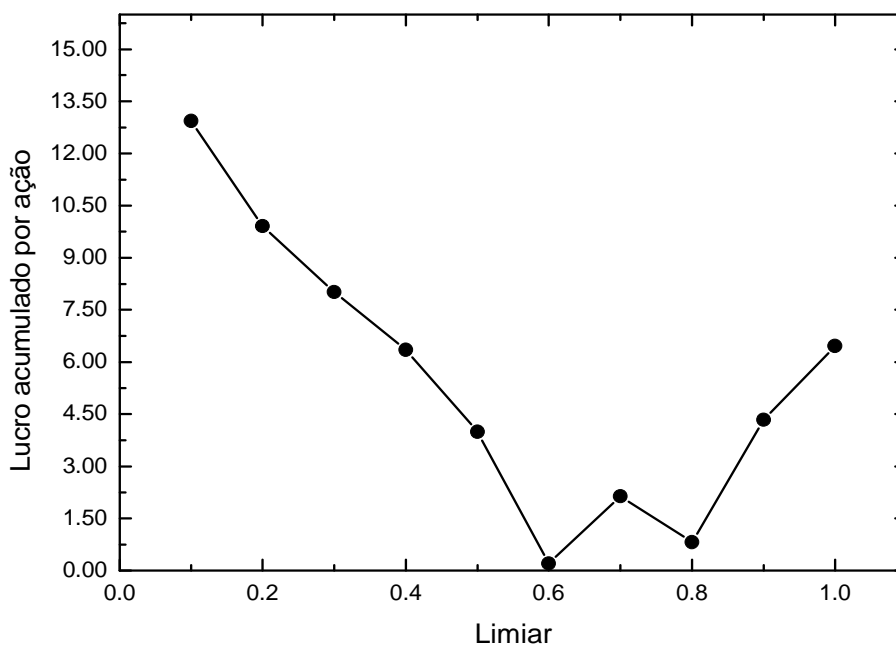


Figura 6.7 Lucro acumulado por ação

O gráfico acima apresenta o lucro (venda – compra) acumulativo por ação (individual) da Vale do Rio Doce para diferentes limiares testados. É possível constatar que não existe uma correlação entre a lucratividade do sistema analisado e os diferentes limiares, ou seja, conforme aumenta o limiar não necessariamente aumenta a lucratividade do sistema. De fato o maior lucro foi obtido com o limiar igual a 0,1% (R\$12,94).

Outro ponto a ser destacado é que o número de operações realizadas (considerando os resultados obtidos com o limiar de maior lucratividade (0,1%)) foi um número bem reduzido (100 operações de compras e vendas em 337 dias de negociação), conforme pode ser observado na figura 6.8. Este é um fato relevante, pois significa que a estratégia adotada foi bem criteriosa em escolher o momento certo de entrar e sair do mercado.

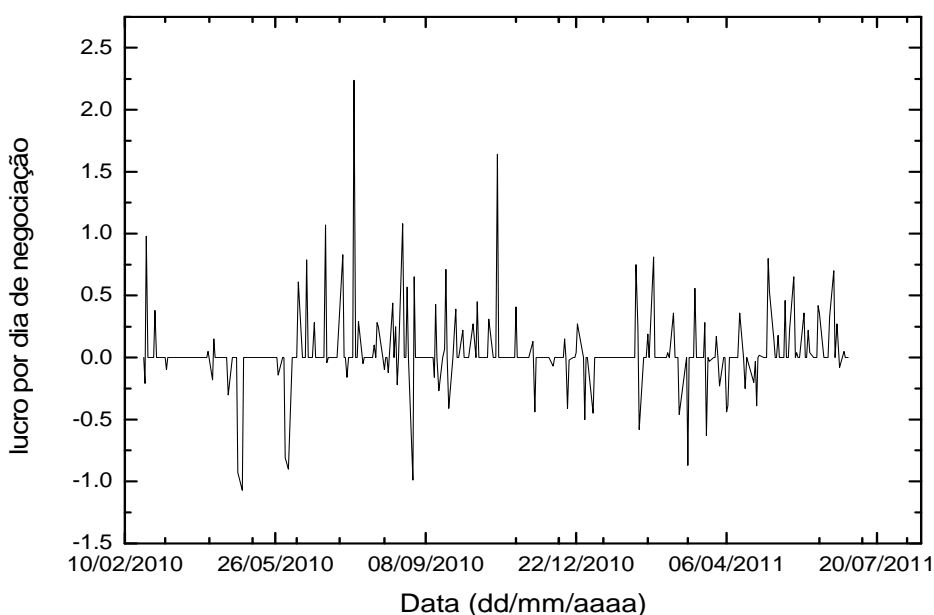


Figura 6.8 Lucro por dia de negociação

Pode-se considerar que resultado obtido (limiar 0,1% e lucro R\$12,94) por ação negociada é um resultado satisfatório, pois compreende a um retorno médio de 33%.

De fato ao comparar os resultados obtidos com a estratégia *Buy and hold* (compra de ações a longo prazo) é possível verificar a eficiência do sistema adotado, pois para o período analisado, o investidor *buy and hold* comprou a ação da Vale na abertura do primeiro dia de negociação (23-02-2010) por um valor de R\$39,42 e vendeu no fechamento do ultimo dia do período analisado (30-06-2011) pelo valor de R\$41,45,

o que equivale a um retorno de aproximadamente 5%, retorno bem menor que o obtido pelo sistema adotado nesta dissertação.

Por fim é importante ressaltar que mesmo com uma acurácia de 66,17% (modelo analisado na estratégia de negociação) foi possível obter um lucro acumulativo satisfatório e maior que o lucro obtido com a estratégia *Buy and Hold*. Isto demonstra que o sistema de previsão desenvolvido conseguiu capturar o movimento das ações da Vale nos dias de maior volatilidade do mercado.

7. Conclusões e Trabalhos Futuros

Neste capítulo uma breve discussão sobre os principais aspectos desta dissertação são apresentados, assim como as principais conclusões e as limitações associadas à implementação do mesmo. Ao final do capítulo, recomendações para trabalhos futuros são apresentados.

7.1 Visão Geral do Trabalho

Estudos envolvendo os vários mercados financeiros mundiais vêm crescendo muito nos últimos anos e atualmente ênfase tem sido dada a utilização de notícias financeiras como entrada para modelos de previsão destes mercados. É amplamente conhecido que notícias podem desempenhar um papel fundamental nos mercados financeiros, entretanto transformar todas estas notícias em conhecimento e em tempo hábil para ajudar os investidores na sua tomada de decisão ainda é um grande desafio para a comunidade científica.

Dentro deste contexto o objetivo central desta dissertação foi prever a tendência diária do principal indicador da Bolsa de Valores de São Paulo (Ibovespa) através da utilização de notícias macroeconômicas e técnicas de mineração de textos. Este objetivo faz parte de uma corrente de trabalhos (muitos deles revisados nesta dissertação) que visam à introdução da metodologia de mineração de textos no estudo dos principais mercados financeiros mundiais.

A metodologia proposta para alcançar os objetivos desta dissertação consistiu de cinco passos (coleta de dados, pré-processamento, classificação das notícias, avaliação dos modelos e estratégia de negociação) que foram implantados através de dois experimentos (Exp1 e Exp2) diferentes. Cada experimento teve seus próprios parâmetros, configurações, enfim suas próprias características, mas a principal diferença entre os mesmos foi a etapa de pré-processamento, onde no Exp1 foi utilizada uma abordagem estatística para a seleção de palavras chave dos modelos, enquanto no Exp2 as palavras chave relevantes do modelo foram fornecidas por um especialista da área do mercado financeiro. Em ambos os experimentos dois algoritmos de classificação diferentes foram utilizados para tentar capturar as relações entre a tendência do

Ibovespa e as notícias utilizadas. Em um passo seguinte diferentes métricas de avaliação foram utilizadas para avaliar os modelos desenvolvidos e por fim uma estratégia de negociação foi proposta para avaliar o modelo de maior acurácia.

7.2 Conclusões gerais

As conclusões apresentadas a seguir estão relacionadas aos resultados obtidos em ambos os experimentos (Exp1 e Exp2) definidos no capítulo 5 e cujos resultados foram analisados no capítulo 6.

Em relação aos parâmetros comuns aos experimentos Exp1 e Exp2 foi possível concluir que:

- A classificação com modelo de apenas duas classes obteve maior acurácia do que os modelos classificados com três classes.
- Base de dados com notícias agrupadas até às 16 horas (Base2) apresentou melhor resultado no Exp1, entretanto para o Exp2 os melhores resultados obtidos para ambas as bases de dados utilizadas (Base1 com notícias agrupadas até às 10 horas e Base2) foram bem semelhantes.

Em relação apenas ao Exp1 podem ser consideradas as seguintes conclusões:

- Utilizando os módulos responsáveis pela atividade de mineração de textos implementados na ferramenta comercial *Poly Analyst* é possível prever o comportamento da tendência diária do Ibovespa com uma acurácia de 59,05%.
- A representação por frequência do documento é mais eficiente que a representação booleana.
- O algoritmo de classificação com maior acurácia para ambas as bases de dados (53,41% para a Base1 e 59,05% para a Base2) foi o SVM.
- As métricas de avaliação utilizadas para avaliar os modelos gerados mostraram que o modelo de maior acurácia possui uma capacidade muito melhor em prever a tendência positiva do que a tendência negativa.

Já em relação ao Exp2 as seguintes conclusões podem ser definidas:

- Não foi possível verificar uma correlação entre os melhores resultados obtidos e a quantidade de neurônios das redes neurais simuladas.
- Nenhuma medida de atribuição de pesos (TF, TF-IDF, TF-CDF) foi eficiente em todas as possíveis configurações de rede analisadas, ou seja, para o modelo de rede MLP e ambas as bases de dados utilizadas (Base1 e Base2) os melhores resultados foram obtidos com a TF-IDF (Base1 = 60,83% e Base2 = 68,55%). Já em relação ao modelo de rede RBF e ambas as bases de dados utilizadas (Base1 e Base2) os melhores resultados foram obtidos com medidas diferentes para cada base de dados, ou seja, para RBF e Base1 foi a medida TF (66,17%) e para RBF e Base2 foi a medida TF-CDF (67,36%).
- Entre os modelos de redes utilizados (MLP e RBF) é possível concluir que nenhum dos dois modelos conseguiu obter os melhores resultados para ambas as bases de dados utilizadas (Base1 e Base2), ou seja, em relação a Base1 o melhor resultado foi obtido com a rede RBF (66,17%), enquanto que para a Base2 o melhor resultado foi obtido com a rede MLP (68,55%).
- A utilização do dicionário de sinônimos criado para este experimento não reportou ter alguma eficiência, uma vez que os resultados com a utilização do mesmo foram inferiores quando comparados aos modelos que não fizeram uso do mesmo.
- Assim como no Exp1, modelo de maior acurácia possui uma capacidade muito melhor em prever a tendência positiva do que a tendência negativa.

Analisando os resultados de uma maneira geral, ou seja, dentro do ponto de vista dos dois experimentos realizados foi possível concluir que os modelos construídos no Exp2 foram muito mais eficientes, de fato o melhor resultado final foi obtido neste experimento (maior acurácia considerada = 66,17%, capacidade para prever a classe positiva = 71,06% e capacidade para prever a classe negativa 59,28%).

Por fim é possível concluir também que os resultados obtidos contradizem a hipótese do mercado eficiente e sugere que através da metodologia proposta nesta dissertação é possível prever o comportamento futuro do mercado de ações com uma rentabilidade considerada.

7.3 Limitações e problemas do trabalho

A maior limitação deste trabalho está relacionada ao escopo de notícias restrito apenas a publicações recentes dos principais jornais econômicos do Brasil, ou seja, os principais sites brasileiros de notícias econômicas não disponibilizam em seus sites arquivos de notícias passadas. Isto não acontece quando se está trabalhando com mineração de dados, pois a própria BmfBovespa disponibiliza séries temporais das cotações de vários ativos de mais de 10 anos atrás.

A utilização de uma ferramenta comercial ajuda a otimizar o tempo de desenvolvimentos dos modelos, pois todos os métodos necessários já estão implementado na ferramenta, entretanto em se tratando da ferramenta utilizada neste trabalho (*Poly Analyst*) a falta de uma documentação mais detalhada sobre a implementação dos diferentes métodos foi um ponto negativo. Outra limitação a ser destacada também é a falta de opção nos métodos utilizados para o pré-processamento dos textos nos módulos relacionados à classificação de textos.

E por fim vale destacar que aplicações envolvendo a utilização de RNAs têm dificuldades em se encontrar uma configuração de parâmetros adequados. Apesar de alguns parâmetros terem sido testados exaustivamente nesta dissertação a falta de um padrão para definir o sucesso das RNAs sempre é um agravante na obtenção de bons resultados. Por este motivo os otimizadores de redes neurais ganharam grande popularidade em aplicações comerciais.

7.4 Sugestões de Trabalhos Futuros

O desenvolvimento de sistemas baseados na análise de notícias e que auxiliem o investidor financeiro em suas tomadas de decisão está ganhando progressiva aceitação dentro da comunidade de investimento (Mitra, 2010), portanto há muitos aspectos que podem ser investigados. No que diz respeito aos diferentes modelos propostos nesta dissertação há muitos problemas que podem ser considerados e investigados em trabalhos futuros.

Uma das questões mais importantes está relacionada aos dados textuais utilizados nesta dissertação (títulos de notícias financeiras). O trabalho proposto confirmou que é possível prever o indicador relacionado à Bolsa de valores de São Paulo com uma acurácia de 66,17% utilizando apenas os títulos das notícias publicados

nos principais jornais de economia do Brasil. Neste sentido um ponto que ficou em aberto é saber se a utilização do título da notícia juntamente com o conteúdo da mesma poderia melhorar a eficiência das classificações e conseqüentemente a lucratividade dos modelos. Sendo assim este ponto é o primeiro passo a ser considerado para aprimorar nossos resultados.

Um segundo ponto a ser considerado em trabalhos futuros é a utilização de um algoritmo de segmentação para extrair as tendências das séries temporais, pois esta estratégia poderia diminuir o ruído presente nas séries, tornando os resultados com a utilização das mesmas mais eficientes.

Outra questão muito discutida ao longo desta dissertação foi qual seria o tempo que o mercado leva para reagir ao conteúdo de uma determinada notícia? Dentro deste contexto, uma abordagem onde diferentes janelas de tempos poderiam ser analisadas ao longo do dia pode ser útil e eficiente. Por exemplo, poderia-se trabalhar apenas com notícias publicadas em um determinado período do dia (uma, duas ou três horas) e fazer uma previsão *intraday* analisando diferentes períodos de publicação de notícias.

Em resumo muitos aspectos podem ser agregados ao método proposto, mas os resultados obtidos são encorajadores e apontam para novas investigações.

REFERÊNCIAS BIBLIOGRÁFICAS

APTE, C., DAMERAU, R., WEISS, S., M., “Automated Learning of Decision Rules for Text Categorization”. **Journal ACM Transactions on Information Systems**, v. 12, Issue 3, pp. 233-251, New York (NY), USA. Jul 1994.

BAKER, L., D., MCCALLUM, M., K., “Distributional clustering of words for text categorization”, In: **Proceedings of SIGIR-98 21st ACM International Conference on Research and Development in Information Retrieval**, pp. 96-103, Melbourne, Australia, 1998.

BASTOS, V., M., **Ambiente de Descoberta de Conhecimento na Web para a Língua Portuguesa**. Tese de D.Sc., COPPE/UFRJ, Rio de Janeiro, RJ, Brasil. 2006.

BMF. Bolsa de Valores de São Paulo e da Bolsa de Mercadorias e Futuros, 2012. Disponível em: <http://www.bovespa.com.br>. Acesso em: 20 jan. 2012.

BOULIS, C., OSTENDORF, M., “Text Classification by Augmenting the Bag-of-Words Representation with Redundancy Compensated Bi-grams”. In: **Proceedings of the International Workshop in Feature Selection in Data Mining**, pp. 9-16, 2005.

BRAGA, A., P., CARVALHO, A., C., P., L., LUDEMIR, T., B., **Redes neurais artificiais: teoria e aplicações**. Rio de Janeiro: LTC, 2000.

CAI, J., SONG, F., “Maximum Entropy Modeling with Feature Selection for Text Categorization”, **Conference of 4th Asia Information Retrieval Symposium (AIRS)**. LNCS, vol. 4993, pp. 549-554, Harbin, China, Jan. 2008.

CASTRO, F., C., C., **Reconhecimento e localização de padrões em Imagens utilizando Redes Neurais Artificiais como Estimador de Correlação Espectral**. Dissertação de M.Sc., PUCRS, Rio Grande do Sul, RS, Brasil. 1995.

CHEN, C.H., “Neural networks for financial market prediction”. In: **Proceedings of the IEEE International Conference on Neural Networks**, v. 2, pp. 1199–1202, Jul, 1994.

CHEN, H., HSU, P., ORWIG, R., HOOPEs, L., NUNAMAKER, J., F., “Automatic Concept Classification of Text from Electronics Meetings”. **Communications of ACM**, v. 37, No. 10, pp. 56-73, Oct. 1994.

CHOU, C., H., SINHA, A., P., ZHAO, H., “A Hybrid Attribute Selection Approach for Text Categorization”, **Journal of the Association for Information Systems (JAIS)**, v. 11, Issue 9, pp. 491-518, Sep, 2010.

CHURCH, K., HANKS, P., “Word association norms, mutual information and lexicography. **Journal Computational Linguistics**, v. 16, No. 1, pp. 22-29, Mar, 1990.

DEBOLE, F., SEBASTIANI, F., “Supervised Term Weighting for Automated Text Categorization”. In: **Proceedings of the 18th ACM Symposium on Applied Computing (SAC)**, pp. 784-788, New York (NY), 2003.

DOAN, S., Horiguchi, S., “An efficient selection using multi-criteria in text categorization”. In: **Proceedings of the 4th International Conference on Hybrid Intelligent Systems (HIS)**, pp. 86-91, Washington, DC, IEEE Computer Society, 2004.

DUMAIS, S., PLATT, J., HECKERMAN, D., SAHAMI, M., “Inductive Learning Algorithms and Representations for Text Categorization”. In: **Proceedings of 7th ACM International Conference of Information and Knowledge Management (CIKM)**, pp. 148-155, New York (NY), Nov.1998.

DUNNING, T., “Accurate Methods for the Statistics of Surprise and Coincidence”, **Journal Computational Linguistics**, v. 19, No.1, pp. 61-74, Mar, 1993.

DZIELINSKI, M., RIEGER, M., O., TALPSEPP, T., “Volatility, asymmetry, news and private investors”. In: Gautan Mitra & Leela Mitra (eds.) **The Handbook of News Analytics for finance**, Chaper 10, New York, USA, John Wiley & Songs, 2010.

EBECKEN, N. F. F., LOPES, M. C. S., COSTA, M. C. A., “Mineração de textos”. In: **Sistemas Inteligentes: Fundamentos e Aplicações**, 1 ed. , Cap. 13, Barueri, SP, Brasil, Manole, pp. 337-370, 2003.

EVERITT, B., S., **Cluster Analysis**. 3rd ed. London: Heinemann Educational Books, 1993.

FAMA, E., F., “Efficient Capital Markets: A Review of Theory and Empirical Work”. **The Journal of Finance** v. 25, n.2, pp. 383-417, May. 1970.

FAMA, E., F., **The Distribution of the Daily Differences of the Logarithms of Stock Prices**. Ph.D. Universidade de Chicago. Chicago, 1964.

FARIA, E., L., ALBUQUERQUE, M., P., GONZALEZ, J., L., CAVALCANTI, J., T., P., ALBUQUERQUE, M., P., “Predicting the Brazilian stock market through neural networks and adaptive exponential smoothing methods”. **Expert System with Applications** v. 36, pp. 12506-12509, May. 2009.

FORMAN, G., “An Extensive Empirical Study of Feature Selection Metrics for Text Classification”, *Journal of Machine Learning*”, **Journal of Machine Learning Research**, v. 3, pp. 1289-1305, 2003.

FRAGOS, K., MAISTROS, Y., SKOURLAS, C., “A weighted Maximum Entropy Language Model for Text Classification”. In: **Proceedings of the 2nd International Workshop on Natural Language Understanding and Cognitive Science (NLUCS)**, pp. 55-67, Miami FL, May, 2005.

FUNG, G., P., C., YU, X., J., LAM, W., “News Sensitive Stock Trend Prediction”. In: **Proceedings 6th Pacific-Asia Conference on Knowledge Discovery and Data Mining**, pp. 481-493, Taipei 2002.

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

GIDÓFALVI, G.: **Using News Articles to Predict Stock Price Movements**. Project Report, Department of Computer Science and Engineering, University of California, San Diego, 2001.

GOADRICH, M., OLIPHANT, L., SHAVLIK, J., “Gleaner: Creating Ensembles of First-Order Clauses to Improve Recall-Precision Curves”, **Journal of Machine Learning**, v. 64, Issue: 1-3, pp. 231-261, Sep. 2006.

HAYKIN, S. **Redes Neurais, Princípios e Prática**. 2.ed. Porto Alegre: Bookman, 2001.

JOACHIMS, T., “A probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization”. In: **Proceedings of the 14th International Conference on Machine Learning (ICML)**, pp.143-151, San Francisco, California: Morgan Kaufmann Publishers Inc., 1997.

JOACHIMS, T., “Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms”. **Journal Computational Linguistics**, v. 29, No. 4, pp. 655-664, Dec, 2003.

JOHN, G., H., KOHAVI, R., PFLEGER, K., “Irrelevant Features and the Subset Selection Problem”. In: **Proceedings of the 11th International Conference in Machine Learning (ICML)**, pp. 121-129, San Francisco, CA: Morgan Kaufmann Publishers, 1994.

KALEV, P., S., LIU, W., PHAM, P., K., JARNECIC, E., “Public information arrival and volatility of intraday stock returns”. **Journal of Banking & Finance** v. 28, pp. 1441-1467, Mar. 2004.

KAUFMAN, L., ROUSSEEUW, P., J., **Finding Groups in Data – An Introduction to Cluster Analysis**. John Wiley and Sons Inc. (Series in Applied Probability and Statistics). New York (NY), 1990.

KEOGH, E., J., CHU, S., HART, D., PAZZANI, M., J., “An Online Algorithm for Segmentation Time Series”. In: **Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM)**. Washington, DC. pp. 289-296, Dec. 2001.

KIM, K., J., “Financial Time Series Forecasting Using Support Vector Machines”, **Neurocomputing** 55, pp. 307-319, Mar 2003.

KIYOSHI I., GOTO, T., MATSUI, T., “Trading Tests of Long-Term Market Forecast by Text Mining”, In: **2010 IEEE International Conference on Data Mining Workshops**, pp. 935-942, Dec 2010.

KROLLNER, B., VANSTONE, B., FINNIE, G., “Stock index forecasting with machine learning techniques: A survey”. In: **Proceedings of the European symposium on artificial neural networks: Computational Intelligence and machine learning**. Bruges, Bélgica, 28-30. Apr. 2010.

KWON, O. W., LEE, J.H., “Web Page Classification Based on K-Nearest Neighbor Approach”. In: **Proceedings of the 5th International Workshop on Information Retrieval with Asian Languages (IRAL)**, pp. 9-15, New York (NY), Sep. 2000.

LAVRENKO, V., SCHMILL, M., LAWRIE D., OGILVIE, P., JENSEN, D., ALLAN, J., “Language Models for Financial News Recommendation”. In: **Proceedings 9th Int. Conference on Information and Knowledge Management**. Washington. pp. 389-396, 2000.

LAVRENKO, V., SCHMILL, M., LAWRIE D., OGILVIE, P., JENSEN, D., ALLAN, J., “Mining of Concurrent Text and Time Series”. In: **Proceedings 6th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining**. Boston. pp. 37-44, Aug. 2000a.

LECUN, Y., JACKEL, L., D., BOTOU, L., CORTES, C., DENKER, J., S., *et al.*, "Learning algorithms for classification: A comparison on handwritten digit recognition". **Neural Networks: The Statistical Mechanics Perspective**, 261-276, 1995.

LEE, L., W., CHEN, S., M., “New Methods for Text Categorization Based on a New Feature Selection Method and a New Similarity Measure Between Documents”, In: **Proceedings of the 19th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AEI)**, v. 4031, pp. 1280-1289, Springer-Verlag, Berlin, 2006.

LEINWEBER, D., J., **Nerds on Wall Street: Math, Machines and Wired Markets**. John Wiley, & Sons, Inc., Hoboken, New Jersey. 2009.

LEWIS, D., D., RINGUETTE, M., “A Comparison of Two Learning Algorithms for Text Categorization”. In: **Proceedings of the 3rd Annual Symposium on Document Analysis and Informational Retrieval (SDAIR)**, pp. 81-93, Las Vegas, Nevada, Apr. 1994.

MAURO, H., **Investimentos: Como administrar melhor seu dinheiro**. 1 ed. São Paulo, Editora Fundamento Educacional, 2001.

MITCHELL, T., M., **Machine Learning**. Boston, USA: McGraw-Hill, 1998.

MITTERMAYER, M., A., “Forecasting Intraday Stock Price Trends with Text Mining Techniques”, **Proceedings of the 37th Annual Hawaii International Conference on System Sciences**, Big Island, pp. 64, Jan 2004.

MITTERMAYER, M., A., KNOLMAYER, G., “Text Mining System for Predicting the Market Response to News: A Survey”, **Working Paper No. 184, Institute of Information Systems, Univ. of Bern**, Bern. Aug 2006.

MITTERMAYER, M., A., KNOLMAYER, G., “NewsCATS: A News Categorization And Trading System”, **Proceedings of the 6th International Conference on Data Mining (ICDM’06)**, Hong Kong 2006a.

MONTANES, E., FERNANDEZ, J., DIAS, I., COMBARRO, E., F., RANILLA, J., “Measures of Rule Quality for Feature Selection in Text Categorization”. In: **Proceedings of the 5th International Symposium on Intelligent Data Analysis (IDA)**, v. 2810, pp. 589-598, Heidelberg, Berlin: Springer-Verlag, 2003.

NIGAN, K., LAFFERTY, J., McCALLUM, A., “A Using Maximum Entropy for Text Classification”. In: **Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI), Workshop on Machine Learning for Information filtering**, pp. 61-67, Stockholm, Sweden, Aug, 1999.

ORENGO, V., M., HUYCK, C., “A Stemming Algorithm for the Portuguese Language”. In: **Proceedings Eighth International Symposium on String Processing and Information Retrieval (SPIRE)**, pp. 186-193, Nov, 2001.

OSUNA, E., FREUND, R., GIROSI, F., “Training support vector machines: An application to face detection”. In: **Proceedings of Computer Vision and Pattern Recognition 97**, pp. 130-136, Jun, 1997.

PAULAK, Z., “Rough sets”, **International Journal of Computing and Information Science**, vol. 11, pp. 341-356, 1982.

PERAMUNETILLEKE, D., WONG, R., K., “Currency Exchange Rate Forecasting from News Headlines”. In: **Proceedings 13th Australasian Database Conference**. Melbourne, Australia. v. 5, pp. 131-139, 2001.

POLY ANALYST, **Poly Analyst Help, Poly Analyst 6** – Megaputer Intelligence Inc. – Tutorial do Software 2007.

REZENDE, S., O., **Sistemas Inteligentes: Fundamentos e Aplicações**, 1 ed. , Barueri, SP, Brasil, Manole, 2003.

ROBERTSON, C., “Enabling Sophisticated Financial Text Mining”. In: **eResearch Australasia**. Novotel Sydney Manly Pacific, Nov. 2009.

ROBERTSON, C., GEVA, S., WOLFF, R., C., “News Aware Volatility Forecasting: Is the Content of News Important?” **Proceedings of 6th Australasian Data Mining Conference (AusDm’07)**. Gold Coast, Australia. v. 70, 2007.

RUMELHART, D. E., HILTON, G., E., WILLIAMS, R., J., “Learning Internal Representations by Error Propagation”. **Parallel Distributed Processing: Exploration in the Microstructure of Cognition**. v.1, MA: MIT Press, Cambridge, 1986. pp. 318-362, 1986.

SALTON, G., BUCKLEY, C., “Term-weighting approaches in automatic text retrieval”, **Information Processing and Management**, v. 24, No. 5, pp. 513-523, Jul, 2002.

SALTON, G., WONG, A., YANG, C., S., “A vector space model for automatic indexing”, **Communications of the ACM**, v. 18, pp. 613 – 620, New York (NY), Nov. 1975.

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

SCHUMAKER, R., P., CHEN, H., “Textual analysis of stock market prediction using financial news articles”, In: **12th Americas Conference on Information Systems (AMCIS-2006)**, Acapulco, México, 2006.

SCHUTZE. H., HULL, D., A., PEDERSEN, J., O., “A comparison of classifiers and document representations for the routing problem”. In: **Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval**, pp. 229-237, Seathe, WA, USA, 1995.

SEBASTIANI, F., “A Tutorial on Automated Text Categorization”. In: **Proceedings of the 1st Argentinian Symposium on Artificial Intelligence (ASAI)**, pp. 7-35, Buenos Aires, Argentina, Sep 1999.

SEBASTIANI, F., “Machine Learning in Automated Text Categorization”, **ACM Computing Surveys**, v. 34, No. 1, pp. 1-47, Mar 2002.

SHANG, W., HUANG, H., ZHU, H., “*et al.*”, “A novel feature selection algorithm for text categorization”, **Expert Systems with Applications**, v. 33, pp. 1-5, 2007.

STRZALKOWSKI, T., “Document representation in natural language text retrieval”. In: **Proceedings of the Workshop on Human Language Technology (HLT)**, pp. 364 – 369, Stroudsburg, PA, USA, 1994.

TETLOCK, C., S., “Giving Content to Investor Sentiment: The role of media in the stock market”. **Journal of Finance** v. 62, pp. 1139-68, May. 2007.

THOMAS, J., D., SYCARA, K., “Integrating Genetic Algorithms and text Learning for Financial Prediction”. In: **Proceedings of the Genetic and Evolutionary Computing Conference (GECCO)**. Las Vegas, Nevada. pp. 72-75, Jul. 2000.

TOKUNAGA, T., IWAYAMA, M., **Text Categorization Based on Weighted Inverse Document Frequency**. Report 94-TR0001. Tokyo, Japan: Department of Computer Science, Tokyo Institute of Technology, Mar, 1994.

TRELEA, I., C., “The Particle Swarm Optimization Algorithm: Convergence analysis and parameter selection”. **Information Processing Letters**, v. 85, pp. 317-325, Sep, 2003.

WANG, Y., WANG, X., J., “A New Approach to Feature selection in Text Classification”. In: **Proceedings of the 4th International Conference on Machine Learning and Cybernetics**, v. 6, pp. 3814-3819, Guangzhou, China, Aug, 2005.

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

WEISS, S. M., INDURKHYA, N., ZHANG, T., DAMERAU, F., **Text Mining - Predictive Methods for Analysing Unstructured Information**, 1 ed. New York, NY, EUA, Springer - Science Business Media, 2005.

WERBOS, P., **Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences**. Ph.D. Thesis, Applied Mathematics, Horward University, Nov. 1974.

WIENER, E., PEDERSEN, J., O., WEIGEND, A., S., “A Neural Network Approach to Topic Spotting. In: **Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR)**, pp. 317-332, Las Vegas, Nevada, Apr. 1995.

WITTEN, I. H., FRANK, E., HALL, M., A., **Data mining: Practical Machine Learning Tools and Techniques**, 3 ed. San Francisco, Springer - Morgan Kaufmann, 2011.

WUTHRICH, B., CHO, V., S., PERAMUNETILLEKE, D., SANKARAN, K., ZHANG, J., LAM, W., “Daily Prediction of Major Stock Indices from Textual WWW Data”. In: **Proceedings 4th ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining**. New York. pp. 364-368, 1998.

YAN, Y., “Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval”. In: **Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval**, pp. 13-22, New York (NY): Springer-Verlag, 1994.

YAN, Y., PEDERSEN, J., O., “A Comparative Study on Feature Selection in Text Categorization”, In: **Proceedings of the 14th International Conference on Machine Learning (ICML)**, pp. 412-420, San Francisco, California: Morgan Kaufmann Publishers., 1997.

YU, L., WANG, S., LAI, K., K., “A Rough-Set-Refined Text Mining Approach for Crude Oil Market Tendency Forecasting”, **International Journal of Knowledge and Systems Sciences**, v. 2, No.1, Mar 2005.

VAPINICK, V., N., **Statistical Learning Theory**. New York (NY): Wiley-Inter Science, 1998.

VAPNIK, V., N., **The Nature of Statistical Learning Theory**. New York: Springer, 1995.

ZAI, Y., HSU, A., HALGAMUGE, S., K., “Combining News and Technical Indicators in Daily Stock Price Trends Prediction”. In: **Proceedings of the 4th International Symposium on Neural Networks (ISNN)**, v. 4493, pp. 1087-1096, Nanjing, China, June 2007.

ZHANG, H., “The Optimality of naïve Bayes”. In: **Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference**, pp. 562-567, 2004.

APÊNDICE A

A tabela apresentada neste apêndice é um resumo sobre os principais trabalhos revisados na literatura e está organizada em quatro seções: objetivo do sistema (onde é fornecida uma ideia do sistema desenvolvido), parâmetros de mineração de textos (onde são apresentados os parâmetros da mineração de textos utilizados nos sistemas desenvolvidos), dados de entrada (são apresentados os dados utilizados) e por último teste (onde é apresentado um resumo dos principais desempenhos reportados pelos autores).

Algumas siglas apresentadas na tabela estão relacionadas aos seguintes conceitos: MP significa modelo proposto e *Randômico* é um simulador onde compras e vendas de ações são simuladas aleatoriamente. Os mercados estudados foram: DJIA e S&P500 (principais índices do mercado dos EUA), NIKKEI (indicador da bolsa de Tóquio), FTSE (índice de Londres), Hang Seng (mercado acionário de *Shangai*), *Straits Times* (Cingapura), e o DAX 100 (mercado acionário alemão). USD-DEM é a taxa de câmbio entre o dólar americano e marco alemão; USD-JPY é a taxa de câmbio entre o dólar USA e o YEN (moeda do Japão). *Roundtrips* são transações de compras e vendas de ações, enquanto *bps* são pontos bases utilizados para calcular os lucros obtidos em uma transação, onde 100 pontos bases significam 1% de lucro.

| Artigos | Wuthrich, 1998 | Lavrenko 2000 | Thomas 2000 | Gidófalvi 2001 | Peramunetille 2002 | Fung 2002 | Mittermayer 2006 |
|-----------------------------------|---------------------|---------------------|-------------------|---------------------|---------------------|---------------------|-------------------------------|
| Objetivo do Sistema | | | | | | | |
| Previsão | Tendência de Preços | Tendência de Preços | Volatilidade | Tendência de Preços | Tendência de Preços | Tendência de Preços | Tendência de Preços |
| Horizonte de previsão | 24 horas | 1 hora | ----- | 1 hora | 3 horas | 1 hora | 15 minutos |
| Parâmetros de Mineração de Textos | | | | | | | |
| Palavras Chave | manual | automático | manual | automatico | manual | automático | Semi-automático |
| Qtd. Palavras Chave | 423 | ----- | 145 | 1000 | 400 | ----- | 85 |
| Classificador | Naive Bayes | Naive Bayes | Regras de Decisão | Naive Bayes | Regras de Decisão | SVM Linear | SVM Polinomial |
| Número de classes | 3 | 5 | 39 | 3 | 3 | 5 (treino:3) | 4 (treino:3) |
| Dados de Entrada | | | | | | | |
| Texto Analisado | Título e corpo | Título e corpo | Título | Título e corpo | Título | Título e corpo | Título e corpo |
| Atribuição de classes | automática | automática | manual | automática | automática | automática | automática |
| Frequência de preços | Fechamento Diário | 10 minutos | Fechamento Diário | 10 minutos | 60 minutos | intraday | 15 segundos |
| Teste | | | | | | | |
| Período investigado | 1997 - 1998 | 1999 - 2000 | 2001 - 2002 | 2001 - 2002 | 1993 | 2002 - 2003 | 2002 |
| Treinamento/Teste | 3 meses | 3 / 1.5 meses | 8 / 5 meses | 5.5 / 2 meses | 1 mês | 6 / 1 mês | Validação cruzada (90% - 10%) |
| MP/Randômico | 44% vs 33% | ----- | ----- | 40% vs 33% | 50% vs 33% | ----- | 45% vs 33% |
| Lucro por roundtrips | 13 bps (0.13%) | 23 bps (0.23%) | 10 bps (0.10%) | 10bps (0,10%) | ----- | ----- | 29 bps (0,29%) |
| Mercados | DJIA, NIKKEI, HS,ST | 127 ações (USA) | Russel 3000 | DJIA | USD/DEM e USD/JPY | 614 ações Hong Kong | SP&500 |

Tabela A.1 – Comparação das principais características dos Sistemas propostos na literatura. (Adaptado de Mittermayer, 2006)