Pontifícia Universidade Católica
do Rio de Janeiro

Heraldo Pimenta Borges Filho

Predição do Comportamento do Mercado Financeiro Utilizando Notícias em Português

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do grau de Mestre em Informática.

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro
Agosto de 2014

Pontifícia Universidade Católica
do Rio de Janeiro

Heraldo Pimenta Borges Filho

Predição do Comportamento do Mercado Financeiro Utilizando Notícias em Português

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

Prof. Ruy Luiz Milidiú
Orientador
Departamento de Informática — PUC–Rio

Prof. Marco Antonio Casanova
Departamento de Informática - PUC-Rio

Prof. Leandro Guimarães Marques Alvim
UFRRJ

Prof. José Eugenio Leal
Coordenador Setorial do Centro
Técnico Científico – PUC-Rio

Rio de Janeiro, 29 de Agosto de 2014

Heraldo Pimenta Borges Filho

Graduou–se em Ciência da Computação pela Universidade Federal Fluminense (UFF). Realizou intercâmbio na Universidade de Coimbra - Portugal, tendo participado de grupos de pesquisa na área de Processamento de Linguagem Natural.

Ficha Catalográfica

Borges Filho, Heraldo Pimenta

Predição do Comportamento do Mercado Financeiro Utilizando Notícias em Português / Heraldo Pimenta Borges Filho; orientador: Ruy Luiz Milidiú. — 2014.

51 f. : il. (color); 30 cm

1. Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2014.

Inclui bibliografia.

1. Informática – Teses. 2. Aprendizado de máquina. 3. SVM. 4. Processamento de Linguagem Natural. 5. Classificaçao de Textos. 6. Mercado de Ações. I. Milidiu, Ruy. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Agradecimentos

A Deus, por tudo.

Aos meus pais e minha querida irmã, pelo exemplo, amor e incondicional apoio, mesmo com minha ausência na vida familiar.

A minha namorada Henriette, pelo seu amor, paciência, compreensão e amizade.

Ao meu orientador, Professor Ruy Milidiú, pelo apoio e incentivo para a realização deste trabalho.

Aos meus colegas do laboratório LEARN da PUC-Rio pelo conhecimento compartilhado e pela amizade.

Aos amigos e professores do CETUC da PUC-Rio, em especial ao professor Marco Grivet e Marcelo Balisteri, pela amizade e apoio técnico.

A todos os colegas, professores e funcionários do Departamento de Informática da PUC-Rio, pelo companheirismo, aprendizado e auxílio.

Pour mon amie Noure-Ayn, pour l'amitié et l'exemple dans un temps très spéciale de notre vie.

Aos meus amigos e professores da Universidade Federal Fluminense, pela amizade, incentivo e formação, em especial minha orientadora, professora Bianca Zadrozny.

Aos meus amigos e professores da Universidade de Coimbra, em especial os Isacs, minha amiga Filipa, aos gajos do laboratório e o professor Paulo Gomes.

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

Resumo

Borges Filho, Heraldo Pimenta; Milidiu, Ruy. **Predição do Comportamento do Mercado Financeiro Utilizando Notícias em Português**. Rio de Janeiro, 2014. 51p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Um conjunto de teorias financeiras, tais como a hipótese do mercado eficiente e a teoria do passeio aleatório, afirma ser impossível prever o futuro do mercado de ações baseado na informação atualmente disponível. Entretanto, pesquisas recentes têm provado o contrário ao constatar uma relação entre o conteúdo de uma notícia corrente e o comportamento de um ativo. Nosso objetivo é projetar e implementar um algoritmo de predição que utiliza notícias jornalísticas sobre empresas de capital aberto para prever o comportamento de ações na bolsa de valores. Utilizamos uma abordagem baseada em aprendizado de máquina para a tarefa de predição do comportamento de um ativo nas posições de alta, baixa ou neutra, utilizando informações quantitativas e qualitativas, como notícias sobre o mercado financeiro. Avaliamos o nosso sistema em um dataset com seis mil notícias e nossos experimentos apresentam uma acurácia de 68.57% para a tarefa.

Palavras–chave

Aprendizado de máquina; SVM; Processamento de Linguagem Natural; Classificaçao de Textos; Mercado de Ações.

Abstract

Borges Filho, Heraldo Pimenta; Milidiu, Ruy (Advisor). **Stock Market Behavior Prediction Using Financial News in Portuguese**. Rio de Janeiro, 2014. 51p. MSc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A set of financial theories, such as the efficient market hypothesis and the theory of random walk, says it is impossible to predict the future of the stock market based on currently available information. However, recent research has proven otherwise by finding a relationship between the content of a news and current behavior of an stock. Our goal is to develop and implement a prediction algorithm that uses financial news about joint-stock company to predict the stock's behavior on the stock exchange. We use an approach based on machine learning for the task of predicting the behavior of an stock in positions of up, down or neutral, using quantitative and qualitative information, such as financial. We evaluate our system on a dataset with six thousand news and our experiments indicate an accuracy of 68.57% for the task.

Keywords

Machine Learning; SVM; Natural Language Processing; Text Classification; Stock Market.

Sumário

Lista de figuras

Lista de tabelas

Foi Ela quem tudo fez.
São João Bosco

1
Introdução

A hipótese do mercado eficiente, introduzida por Fama em 1965 [1], afirma que o mercado financeiro seria informacionalmente eficiente. Dessa forma, não seria possível desenvolver um sistema preditor baseado em informações disponíveis, pois esta já estaria refletida no atual preço das ações. No mesmo sentido, em 1985, [2] argumenta, pela teoria do passeio aleatório, que o preço dos ativos evolui em um padrão comparado a um passeio aleatório, ou seja, nenhum sistema de predição seria melhor que um sistema puramente aleatório [3]. Evidências demonstram que tais teorias não são capazes de explicar fenômenos regularmente observados no mercado financeiro [3]. Pesquisas recentes apresentam uma forte relação entre a hora da liberação de uma notícia e a reação do mercado a partir dessa publicação [4]. Vários trabalhos foram produzidos na tentativa de predizer o movimento do mercado financeiro baseado no conteúdo de novas notícias [3, 4, 5, 6, 7, 8, 9], demonstrando a dificuldade da tarefa.

1.1
Problema

Neste trabalho, investigamos o impacto da publicação de notícias jornalísticas, escritas em português, sobre o mercado financeiro brasileiro. Nosso objetivo é construir um modelo que prediz o comportamento de um ativo nas posições de alta, baixa ou neutra utilizando a combinação de informações quantitativas como o histórico dos preços, e informações não quantitativas, como as notícias. Com esse modelo, é possível criar um sistema que indique o comportamento de um ativo de determinada empresa após a publicação de uma notícia relacionada a essa mesma empresa.

1.2
Metodologia

A fim de investigar a influência de notícias sobre o movimento do preço de ativos, implementamos técnicas de aprendizado de máquina e de processamento de linguagem natural para a criação de um modelo de predição.

Com a aplicação dessas técnicas, é estabelecida a relação entre os dados numéricos e os dados textuais. Adicionalmente, um modelo de predição pode ser aprendido através da tarefa de classificação de texto. Cada nova notícia é classificada dentre a possibilidade de três classes, podendo ser considerada "alta", "baixa" ou "neutra", indicando respectivamente o comportamento de alta, baixa, ou de variação não significativa do preço de um ativo. Dessa forma, uma notícia classificada como "neutra" indica que o ativo irá se comportar com pequena ou nenhuma variação dentro de um horizonte de predição. Caso a notícia seja classificada como "alta", indica uma variação positiva e, analogamente, quando classificada como "baixa", trata-se de uma variação negativa.

Para a resolução da tarefa de classificação, criamos um *dataset* com 19.997 noticias sobre o mercado financeiro em relação à Petrobras, companhia petrolífera brasileira. Adicionalmente, criamos uma base de dados com informações numéricas *intraday* de ativos da BMF&Bovespa. A primeira etapa de nossa abordagem é extrair atributos de cada uma das notícias. Em seguida, anotar automaticamente um subconjunto das notícias entre as classes citadas. Finalmente, utilizamos os atributos gerados e o subconjunto de noticias anotadas para construir um modelo SVM. Como métrica de avaliação adotamos a acurácia que mede a percentagem das notícias classificadas corretamente. Nossos resultados experimentais indicam uma acurácia de 68,57%.

1.3 Organização

Este trabalho está organizado da seguinte forma. No Capítulo 2, investigamos os principais trabalhos que utilizam notícias financeiras para predição do mercado. No capítulo 3, descrevemos a tarefa de classificação de textos e os critérios de avaliação. No capítulo 4, descrevemos a metodologia utilizada para a modelagem do problema e criação dos experimentos. No Capítulo 5, descrevemos os experimentos realizados e nossos resultados, comparando-os com o *baseline*. Finalmente, no capítulo 6, nossas conclusões são apresentadas.

2
Revisão da Literatura

Diversas pesquisas têm sido desenvolvidas utilizando técnicas de aprendizado de máquina para a construção de preditores no mercado financeiro. Na maioria dos trabalhos, os preditores são construídos apenas com informações quantitativas, alimentados com atributos como histórico de preço, volume e indicadores de análise técnica. Outros trabalhos, em menor número, utilizam informações qualitativas, como notícias sobre o mercado financeiro [3].

Em 1998, Wuthrich [4], empregando algoritmos de aprendizado de máquina como *Naïve Bayes*, *Nearest Neighbor* e Redes Neurais, apresentou um preditor diário que utilizava apenas o léxico das notícias.

Dois anos depois, em 2000, Lavrenko [5] introduziu um preditor que utilizava a técnica *Term Frequency - Inverse Document Frequency* (tf-idf) [10] como seletor de atributos. Para categorizar as notícias em umas das cinco categorias por ele utilizadas, o autor fez uso de Regressão Linear. Já na etapa de classificação, o algoritmo utilizado foi o *Naïve Bayes*.

Em 2001, Gidófalvi [6] apresentou evidências de que notícias influenciam o mercado dentro de uma janela de tempo de 20 minutos. Com isso, utilizou uma heurística que verificava o aumento e a redução do preço de uma ação para categorizar a notícia em uma das três categorias propostas, ao invés de utilizar Regressão Linear como Lavrenko [5]. Para a extração de atributos, foi utilizada a técnica *Bag-of-Words*, e para a seleção de atributos a técnica de *Mutual Informations*(MI). Assim como em Lavrenko [5], fez uso do algoritmo *Naïve Bayes* para a classificação de notícias.

No ano de 2002, Peramunetilleke [7] representou as notícias através de um dicionário, extraindo os atributos de regras geradas manualmente. Nesse trabalho, o autor utilizou uma janela de predição que variava de uma a três horas e obteve uma acurácia de 50% na classificação das notícias em uma das três classes propostas utilizando o classificador *Naïve Bayes*.

Um sistema desenvolvido por Mittermayer [8] em 2006, NewsCats, classificava as notícias em três categorias: *GOOD, BAD, NEUTRAL*. Para realizar a anotação, o sistema analisava a variação de preços médios em uma janela de tempo de 15 minutos a partir da liberação da notícia. Para a representação do

texto, utiliza *Bag-of-Words*, com métodos de seleção de atributos como IDF e CTF. Nesse caso, o algoritmo de classificação utilizado que obteve melhor resultado foi o SVM, com Kernel polinominal.

Utilizando o classificador SVM para ações na Hong Kong Stock Exchange, Li [9], em 2011, obteve uma acurácia de 64,23%. Para anotar as notícias em uma das três categorias sugeridas o autor utilizou Regressão Linear.

3
Classificação de Textos

O interesse pela tarefa de classificação de textos tem crescido nos últimos anos devido a maior disponibilidade de documentos em formato digital e a consequente necessidade de organizá-los. Com o aumento do volume de informações disponíveis na Internet e seu crescimento exponencial, há uma necessidade cada vez maior de ferramentas que ajudem usuários a filtrar e gerenciar tais dados. Neste capítulo apresentamos os algoritmos e técnicas adotadas ao longo da dissertação e sua utilização para a geração do modelo.

3.1
Visão Geral

Podemos entender classificação como a tarefa de determinar se um objeto o_i, pertencente a um conjunto de objetos O, pertence ou não a uma classe c_j, dentro de um conjunto de classes C. Esta tarefa era tipicamente realizada por especialistas do domínio de determinado problema. Executada por humanos, se torna bastante demorada e custosa, tendendo a ser impraticável executá-la em grandes volumes de documentos. Com o advento das grandes bases de dados como as de portais de notícias e bibliotecas digitais, além de seu uso massivo, técnicas de classificação automática, realizada por algoritmos, substituem a classificação manual. Dessa forma, tornam-se peça chave para a organização e gerenciamento de grande quantidade de informação e essenciais para o desenvolvimento de novas tecnologias e aplicações para usuários.

A abordagem dominante para o problema de classificação automática de textos é baseada em técnicas de aprendizado de máquina, dentro de um processo indutivo criado a partir de um conjunto de objetos pré-classificados em classes definidas [10]. Nessa abordagem, um classificador será automaticamente construído, devendo ser capaz de classificar novos documentos, não utilizados na etapa de aprendizagem. Essa abordagem tem inúmeras vantagens sobre a classificação manual, pois, além de ser economicamente mais barata, tem uma excelente taxa de precisão e pode ser adaptada para diferentes domínios. Diversos métodos baseados em teoria estatística e aprendizado de máquina tem sido adaptados e aplicados nos últimos anos na solução de problemas de classi-

ficação [11], podendo destacar algoritmos de aprendizado *Bayesiano* e *Support Vector Machine* [12].

3.2 Aplicações

A ideia de classificação é universal, podendo ser utilizada em diversas aplicações do mundo comercial. Classificação de textos é uma área importante de pesquisa com diversas aplicações dentro da área de Recuperação de Informação [10]. Um classificador pode ser aplicado em diversos domínios, como marketing direcionado, organização de documentos e diagnóstico médico. Podemos citar ainda outros exemplos relevantes, tais como:

- Identificação do idioma de um site. O resultado de um motor de busca pode ser otimizado identificando o idioma dos sites nos quais se realiza uma determinada pesquisa. Dessa forma, uma busca personalizada seria realizada para o idioma do usuário ou outro idioma de sua escolha;

- Análise de Sentimentos. Esta área de pesquisa tenta identificar o sentimento expresso por usuários a respeito de alguma entidade de interesse, tal como um filme, uma pessoa ou um produto [13]. Nesta tarefa, uma opinião pode ser classificada como positiva ou negativa em relação à informação que é apresentada. Um exemplo de sua aplicação é a detecção de opiniões negativas sobre determinado produto. Esta informação poderá auxiliar um consumidor na decisão de compra do referido produto;

- Triagem de E-mails. Um classificador automático poderá ajudar um usuário a organizar seus e-mails, os associando a categorias de assuntos específicos, como pessoal, financeiro e trabalho. Uma aplicação comum desta tarefa é a detecção de *spams*. Os e-mails recebidos pelo usuário são classificados em spam e não spam, dessa forma, e-mails úteis e não úteis para o usuário;

- Identificação de tópicos. Um classificador automático pode ser utilizado na tarefa de filtrar notícias de jornais *online* que sejam de interesse de um usuário, criando um ambiente personificado, adequado a suas necessidades.

A seguir serão apresentados conceitos sobre a tarefa e o processo de classificação. A seção 3.3 traz a definição formal do problema e descreve sua

aplicação na tarefa de classificação de textos.A seção seguinte apresenta o processo de classificação e cada uma de suas etapas e os critérios adotados. Estas etapas nortearam as decisões adotadas durante a modelagem do problema. A seção 3.4.6 apresenta dois dos principais algoritmos de aprendizagem utilizados para classificação de texto: *Naïve Bayes* e *SVM*. Por fim, apresentamos considerações finais que concluem este capítulo.

3.3 Definição do Problema

Classificação é a tarefa de atribuir objetos em classes previamente definidas [14]. Uma classe é um conjunto de objetos cujo conteúdo pode ser descrito pelo seu rótulo. Neste trabalho, utilizamos uma modelagem preditiva, na qual o modelo de classificação é utilizado para prever o rótulo de objetos ainda não classificados. Técnicas de classificação são mais apropriadas para classes nominais ou binárias, por não considerar a ordem implícita entre as classes. Esse modelo pode ser interpretado como uma caixa preta, que, após ser modelado, atribui uma ou mais classes a um novo documento.

Uma das diferentes aplicações da tarefa é a classificação de documentos, ou mesmo de textos. Seu objetivo é identificar e classificar o tópico ou o assunto de um documento. Considere, por exemplo, um conjunto de notícias relacionadas a uma determinada empresa que gostaríamos de dividir entre boas e ruins. Este exemplo pode ser modelado como um problema de classificação binária, com as classes "notícias boas"e "notícias ruins". Cada classe se refere ao valor subjetivo da notícia, se positivo ou negativo em relação à empresa.

Podemos definir formalmente o problema da classificação textual como a tarefa de aprender uma função alvo F, conhecida como modelo de classificação, que, dado uma coleção D de documentos e um conjunto $C = \{c_1, c_2, ..., c_L\}$ de L classes, atribui a cada par $[d_i, c_j]$ o valor 0 ou 1. Se o valor atribuído for 1, entende-se que o documento d_i pertence a classe c_j, caso contrário, d_i não pertence a c_j [10].

Pela definição, o mesmo documento pode ser atribuído a mais de uma classe. Dessa forma o classificador seria do tipo multi-rótulo. Caso haja a restrição para a classificação em apenas uma classe, dizemos que o classificador é do tipo único-rótulo. Muitas vezes, problemas de multi-rótulo podem ser transformados em problemas de único-rótulo. Dado o problema abordado neste trabalho, utilizamos um classificador que atribui a uma notícia apenas uma única classe. Logo, além de verificar a pertinência de um documento em cada classe, é necessário decidir qual a melhor classe para este documento [15].

3.4 Processo de Classificação

O processo de classificação de textos envolve uma série de etapas. Baeza [10] afirma que para realizar a tarefa é necessário implementar seis etapas, dividindo o processo em coleta de documentos, pré-processamento, representação dos documentos, seleção de atributos, indução de um classificador e avaliação do desempenho, como apresentado em 3.1:

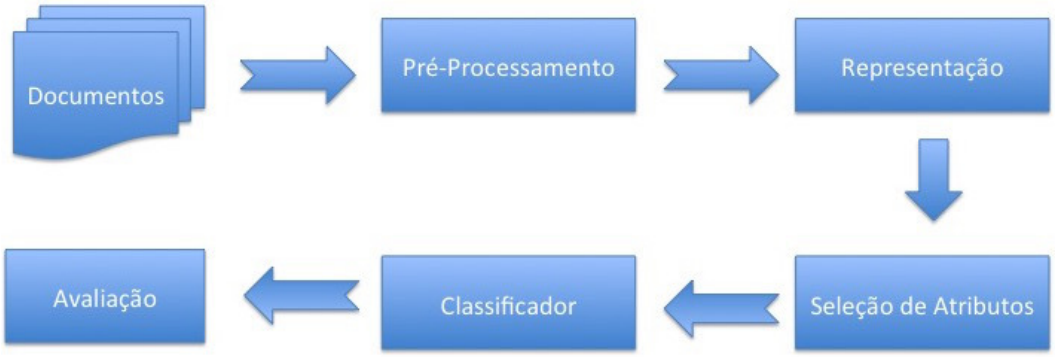


Figura 3.1: Processo de classificação de textos

3.4.1 Coleta de documentos

O primeiro passo para realizar a classificação de textos é coletar os dados que serão utilizados para a tarefa de treinamento. Nesta etapa são selecionados os documentos pertencentes ao domínio do problema que serão classificados. Esses documentos podem ser fornecidos ou obtidos através de bases de dados. Outra grande fonte de documentos que deve ser considerada é a internet, através das coleções de páginas web.

Para selecionar os documentos relevantes para o trabalho, é necessária a criação de algum filtro de seleção. Através desse filtro, documentos não relacionados ao domínio do problema são descartados. Dado as diferentes fontes de informação e a grande variedade de formatos digitais, é importante converter os documentos para um formato padrão, que possibilite as operações de extração de atributos dos textos. Dessa forma, o modelo a ser implementado pode ser aplicado a todos os documentos de forma uniforme. Um padrão que pode ser adotado é o formato XML (*Extensible Markup Language*). O padrão XML especifica uma linguagem que possibilita a fácil identificação da estrutura do texto através do uso de rótulos.

3.4.2
Pré-processamento do texto

Esta etapa engloba as tarefas de remoção de *stopwords*, termos que ocorrem com grande frequência, a tarefa de tokenização, onde é realizada a quebra do texto em unidades menores, e *stemming* de palavras, onde uma palavra é substituída por sua forma canônica.

– Tokenização. O processo de tokenização é utilizado para decompor o documento a ser analisado em termos menores, possibilitando a separação dos elementos constituintes do texto. Em geral, o processo de separação ocorre com a quebra do texto em palavras. Os delimitadores utilizados para tokenização geralmente são: o espaço em branco entre os termos, quebras de linhas, tabulações, e alguns caracteres especiais;

– Stopwords. *Stopwords* são palvras que não apresentam informação significante no contexto do texto em que se encontram, dessa forma, palavras consideradas irrelevantes para a análise do texto. Em geral, este conjunto de palavras é formado por artigos, preposições e pronomes, e ainda termos que aparecem com alta frequência. Nesse processo, os termos ditos *stopwords* são descartados do texto.

– Stemming. Este processo é realizado para cada palavra de forma isolada. Tem por objetivo remover o prefixo e o sufixo dos termos que podem apresentar variação verbal, plural e ainda de gênero. Essa tarefa auxilia a tarefa de classificação, pois reduz o número de termos utilizados para a representação do documento;

3.4.3
Representação do Documento

Nesta etapa os atributos que serão utilizados para representar o texto são definidos. Para cada atributo, deve ser associado um peso de acordo com sua relevância.

3.4.4 Seleção de atributos

Um dos grandes problemas da tarefa de classificação é a alta dimensão do espaço de atributos. A grande ideia da etapa de seleção de atributos é selecionar um subconjunto de características do documento original conforme um critério de seleção, reduzindo a dimensionalidade do problema. Essa seleção melhora a eficiência, a escalabilidade e precisão do classificador.

3.4.5 Construção do Modelo

Um classificador é gerado a partir das representações ponderadas dos documentos do conjunto de treino. Nas seções seguintes abordaremos dois dos algoritmos mais difundidos na literatura, o *Naive Bayes* e o *SVM*, este último utilizado na modelagem deste trabalho.

3.4.6 Aprendizado de Máquina

Segundo Mitchell [12], Aprendizado de Máquina(AM) é uma área preocupada com o projeto e o desenvolvimento de algoritmos que aprendem padrões dos dados fornecidos como entrada. Os padrões aprendidos poderão ser utilizados para realizar predições de novos dados. AM é uma subárea de pesquisa da Inteligência Artificial [16] e engloba os estudos dos métodos computacionais para a automação da aquisição de conhecimento e para a estruturação e acesso do conhecimento já existente.

Essa área é multidisciplinar, podendo ser aplicada em diferentes outras áreas. De acordo com Mitchell [12], algoritmos de AM tem demonstrado serem de grande valor prático para uma variedade de domínio de aplicações. São especialmente úteis em problemas como Mineração de Dados, onde é feita a análise de grandes bancos de dados; em domínios dinâmicos, onde o sistema necessita adaptar-se dinamicamente a mudanças; em processamento de imagens, por exemplo, na tarefa de reconhecimento facial; na análise de mercado de ações; e em problemas de processamento de linguagem natural(PLN) [17].

Os algoritmos de AM são dependentes da etapa de aprendizado, a qual detecta padrões nos dados de entrada e os utiliza para a geração de um modelo. Dependendo do mecanismo de aprendizado, tais algoritmos podem ser basicamente classificados como de aprendizado supervisionado e não supervisionado.

O *aprendizado não supervisionado* utiliza como dado de entrada exemplos não rotulados. Dessa forma, é preciso identificar como agrupar objetos em conjunto desconhecido de classes, através da proximidade dos atributos dos

mesmos. São exemplos de algoritmos não supervisionados as redes neurais e as técnicas de *clustering*.

Por sua vez, o *aprendizado supervisionado* distingue-se do *aprendizado não supervisionado* por receber como entrada dados de treinamento. Estes dados são compostos por pares de atributos-classe, documentos pré-classificados por seres humanos, os quais serão utilizados para treinar o classificador. Na literatura, encontramos diversos algoritmos para aprendizado supervisionado, cada qual utilizando uma técnica diferente para determinar a função aprendizado. A seguir, apresentaremos os algoritmos supervisionados para a classificação de textos abordados neste trabalho.

3.4.7 Naïve Bayes

Os classificadores probabilísticos atribuem a cada par documento-classe $[d_j, c_p]$ uma probabilidade de que o documento pertença a uma determinada classe. Após computar a probabilidade de cada par, o classificador atribui ao documento d_j as classes com maior probabilidade. Para calcular a probabilidade de um documento pertencer a uma classe, um classificador probabilístico aplica o teorema de *Bayes*, como se segue:

$$P(classe|atributos) = \frac{P(classe) \cdot P(atributos|classe)}{P(atributos)}$$

Um classificador de *Bayes* simples utiliza uma simplificação que consiste em assumir independência entre os atributos que representam o documento. Dado que essa simplificação não é real, esse classificador é conhecido como *Bayes Ingênuo*, ou mesmo *Naïve Bayes*.

A suposição da independência entre os atributos de um documento pode ser formalmente declarada como:

$$P(classe|atributos) = \frac{P(classe) \cdot P(a_1|classe) \cdot \ldots \cdot P(a_n|classe)}{P(atributos)}$$

Ao invés de calcular a probabilidade de $P(atributos)$, o algoritmo calcula apenas o denominador para cada classe, normalizando de forma que a soma seja sempre 1. Assim este termo, que chamamos de evidência, é uma constante que pode ser ignorada no momento de decidir qual classe tem maior probabilidade.

$$P(classe|atributos) = \frac{P(classe) \cdot P(a_1|classe) \cdot \ldots \cdot P(a_n|classe)}{evidência}$$

Uma vez que todas as probabilidades dos pares documento-classe tenham sido computados, o classificador atribui a cada documento a classe com maior pontuação. Para este trabalho, utilizamos a implementação do classificador *Naïve Bayes* fornecido pela biblioteca *Scikit*, desenvolvida na linguagem *python* [18].

3.4.8 Support Vector Machine

O classificador SVM (*Support Vector Machine*) é um algoritmo de aprendizado de máquina proposto inicialmente por Vapnik em 1992 [19] para tarefas de identificação de padrões com diversas vantagens na resolução de problemas que possuam pequenas amostras e grande número de dimensões. Apesar de trabalhar apenas com dados numéricos, em 1998, foi utilizado pela primeira vez para a tarefa de classificação de textos [1] e tem se apresentado como o mais adequado para esta tarefa [20].

O SVM trabalha essencialmente com uma abordagem geométrica para o problema de classificação binária. Os atributos dos documentos compõe um espaço m-dimensional no qual cada documento é representado por um vetor. A proposta é encontrar uma superfície de decisão, um hiperplano, que será utilizada para separar os novos documentos em duas classes, c_a e c_b. O hiperplano, que é gerado a partir dos exemplos do conjunto de treinamento, divide o espaço em duas regiões de forma que os documentos da classe c_a estejam em uma região e os da classe c_b em outra região. Após a criação do hiperplano, um novo documento d_j será classificado em uma das classes c_a ou c_b pela sua posição relativa ao hiperplano separador.

A teoria sobre *Support Vector Machine* é baseada na ideia de maximizar a distância entre as classes, encontrando o melhor hiperplano separador[21]. Considere um exemplo bidimensional cujos vetores dos exemplos de treinamento sejam linearmente separáveis, como ilustrado na figura 3.2. De todas as linhas que separam os documentos nas duas classes, a linha s maximiza as distâncias para os documentos mais próximos de cada classe, dessa forma constitui o melhor hiperplano separador[10]. A linha r, apesar de prover uma alternativa, tem sua distância menor aos exemplos das classes c_a e c_b. Ao hiperplano de melhor separação é dado o nome de *hiperplano de decisão*. As linhas paralelas pontilhadas, como ilustra o exemplo da figura 3.2, são conhecidas por *hiperplanos delimitadores* e os documentos que pertencem a esses hiperplanos são conhecidos como *vetor de suporte*.

Uma limitação do *SVM* é trabalhar apenas com atributos numéricos. É necessário converter as notícias em vetores, de forma a fornecer o máximo de

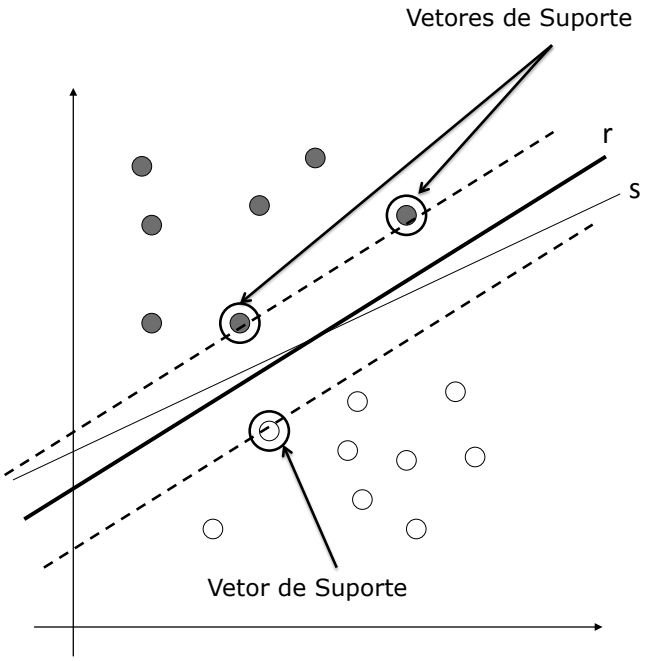


Figura 3.2: Hiperplano com a Máxima Distância

informações. Assim, para representar a notícia em vetores, a cada atributo é associado um índice numérico sequencial e para coordenada do vetor é atribuído o valor 1, caso o atributo esteja presente, ou 0, em sua ausência. Este tipo de representação é conhecido como representação por presença. Nos experimentos realizados, mostrou-se mais eficiente que a representação por frequência, na qual é atribuído a coordenada do vetor a frequência de um termo.

Para este trabalho usamos a implementação do *SVM* conhecida como LibLinear[22], que utiliza a função kernel linear. Essa implementação é uma versão otimizada para tarefas onde o número de atributos é maior que o número de exemplos de treino.

3.4.9 Avaliação do Desempenho

Nesta etapa a classificação realizada é avaliada, utilizando um conjunto de teste. São usadas algumas métricas para quantificar as taxas de erro e de acerto. A avaliação é um passo muito importante para o desenvolvimento de um método de classificação de texto. Sem a avaliação adequada, não há como determinar a qualidade do classificador proposto. Para avaliar o desempenho dos modelos gerados, nós empregamos as métricas usuais da literatura. Na

figura 3.3 é apresentada uma matriz de confusão e em seguida descrevemos as métricas utilizadas.

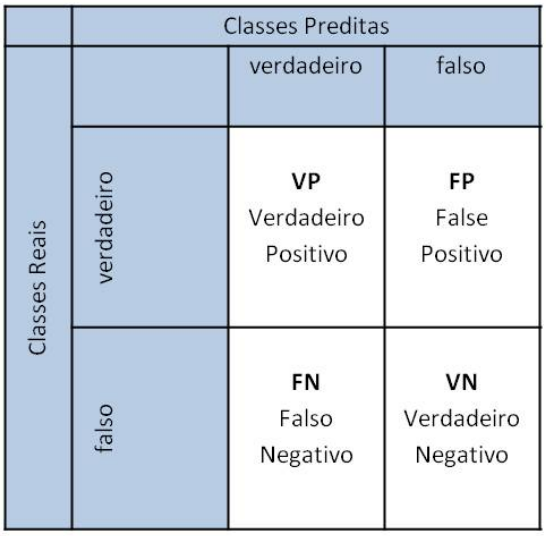


Figura 3.3: Matriz de Confusão

Acurácia

Mede a percentagem dos dados preditos que estão corretos. A acurácia é a fração dos documentos de treinamento atribuídos a suas classes corretas pelo classificador. É representada pela seguinte fórmula:

$$acuracia = \frac{VP + VN}{VP + VN + FP + FN}$$

O problema dessa métrica é que ela não reflete a realidade dos resultados quando as classes estão desbalanceadas.

Precisão

Precisão é a fração de todos os documentos atribuídos à classe c_p pelo classificador, que realmente pertence a classe c_p. É representada pela seguinte fórmula:

$$precisao = \frac{VP}{VP + FN}$$

Recall

Recall é a fração de todos os documentos que pertencem à classe c_p que foram corretamente atribuídos a classe c_p pelo classificador. É representada pela seguinte fórmula:

$$acuracia = \frac{VP + VN}{VP + VN + FP + FN}$$

O problema dessa métrica é que ela não reflete a realidade dos resultados quando as classes estão desbalanceadas.

F1

É uma média harmônica entre os valores de *recall* e precisão. É definida pela seguinte fórmula:

$$F_{\beta=1} = \frac{precisao.recall}{precisao + recall}$$

3.4.10 Metodologia de Treino e Teste

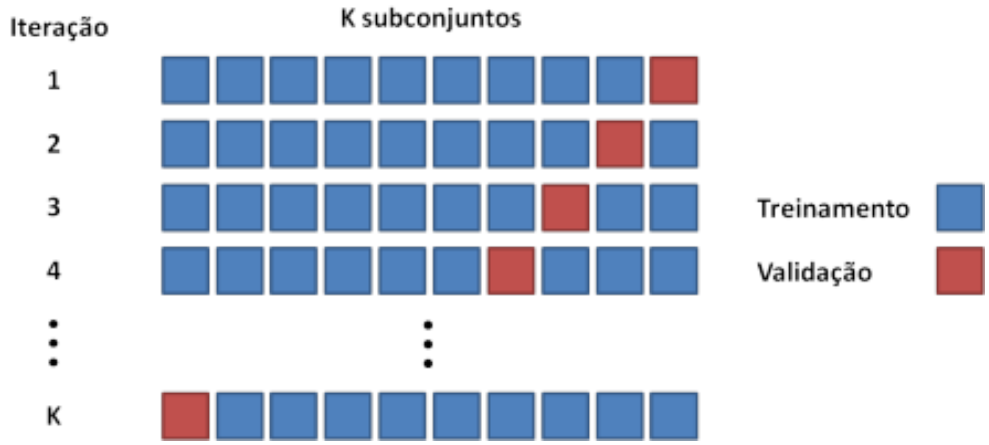


Figura 3.4: Validação Cruzada para k Partições

A validação cruzada é um método que garante a validação estatística dos resultados da classificação. Este método consiste na divisão em k partes disjuntas da amostra original, como exemplificado na figura 3.4, onde $k - 1$ partes são utilizados para a etapa de treinamento e 1 parte é utilizada para a etapa de teste. Este processo é repetido k vezes, alternando as partes sem repetição, de forma que todas as partes sejam utilizadas como treino e teste. Em cada repetição, os conjuntos de treino e teste são avaliados por alguma métrica de avaliação.

3.5 Considerações Finais

Neste capítulo abordamos os conceitos básicos para o entendimento deste trabalho. Nas duas primeiras seções, apresentamos uma visão geral do pro-

blema de classificação de textos e exemplos de sua aplicação em problemas relacionados. Na seção seguinte, vimos a definição formal do problema de classificação. Por fim, na última seção, apresentamos o processo de classificação, elucidando o uso de algoritmos de aprendizado de máquina, entre estes o *Support Vector Machine*. Concluímos que o problema abordado neste trabalho pode ser resolvido através da tarefa de classificação de textos. No capítulo 4, a seguir, veremos como utilizar a tarefa de classificação para resolver o problema de predição do mercado financeiro utilizando notícias em português e a modelagem adotada.

4
Metodologia

Um sistema de predição baseado em notícias pode ser modelado como uma tarefa de classificação de textos. A meta é prever algum aspecto do mercado de ações, como preço, volatilidade ou comportamento baseado no conteúdo das notícias. Um sistema de predição pode ser dividido em duas fases principais: a fase de treino e a fase operacional. Na fase de treino, os dados de entrada são utilizados para modelar o classificador. Na fase operacional, o classificador irá atribuir a uma nova notícia umas das classes definidas.

A abordagem adotada neste trabalho é baseada na classificação de notícias nas classes de "alta", "baixa"ou "neutra". A anotação de uma notícia na classe "alta"indica uma tendência de elevação de preços, na classe "baixa"uma tendência a queda dos preços e se na classe "neutra"indica um comportamento com pequena ou nenhuma variação.

Um modelo de predição foi implementado a fim de atingir os objetivos propostos neste trabalho e é descrito neste capítulo.

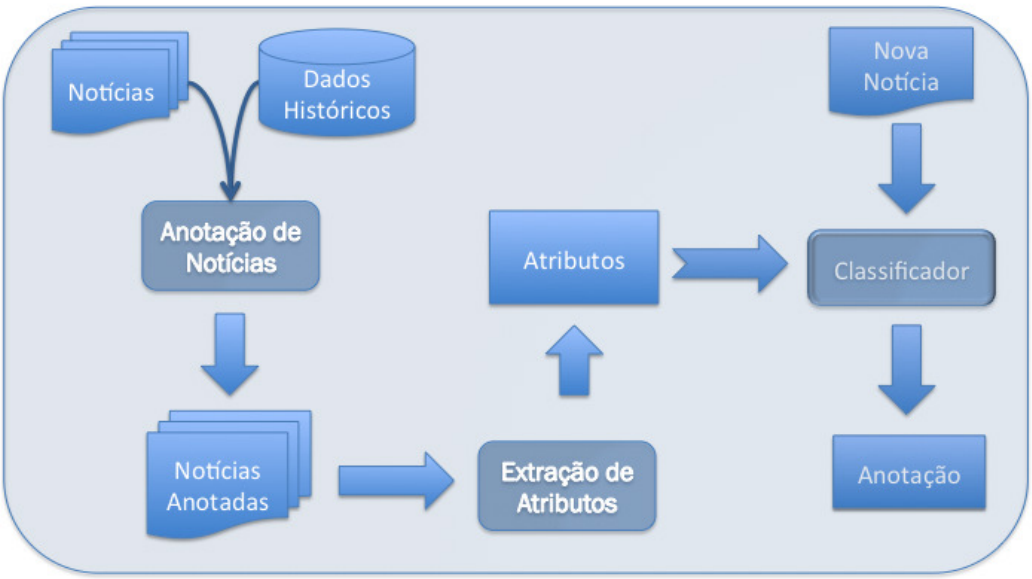


Figura 4.1: Sistema de Predição

4.1 Processo de Predição

O sistema implementado consiste de uma série de etapas, dentro do processo de predição. Em primeiro lugar, foram criados os *dataset* de notícias e de dados históricos do mercado. Em seguida, essas notícias foram anotadas utilizando os dados numéricos das séries históricas em um etapa de anotação automática. No passo seguinte, os atributos do texto foram extraídos para serem então utilizados para treinamento e teste do classificador. Todo esse processo é apresentado na figura 4.1. Este trabalho foi implementado utilizando a linguagem de programação *Python*, na versão 2.7.

4.2 Coleta de Dados

Nesta seção são descritas as tarefas realizadas para a coleta e organização dos dados necessários para a execução deste trabalho. Dessa forma, foram criadas duas grandes bases, sendo uma com informações textuais, compostas com notícias sobre a Petrobras e uma base com dados numéricos, extraídos das movimentações do mercado financeiro.

4.2.1 Dataset de Notícias

Diversos *Web Sites* fornecem notícias em tempo real sobre o mercado financeiro.

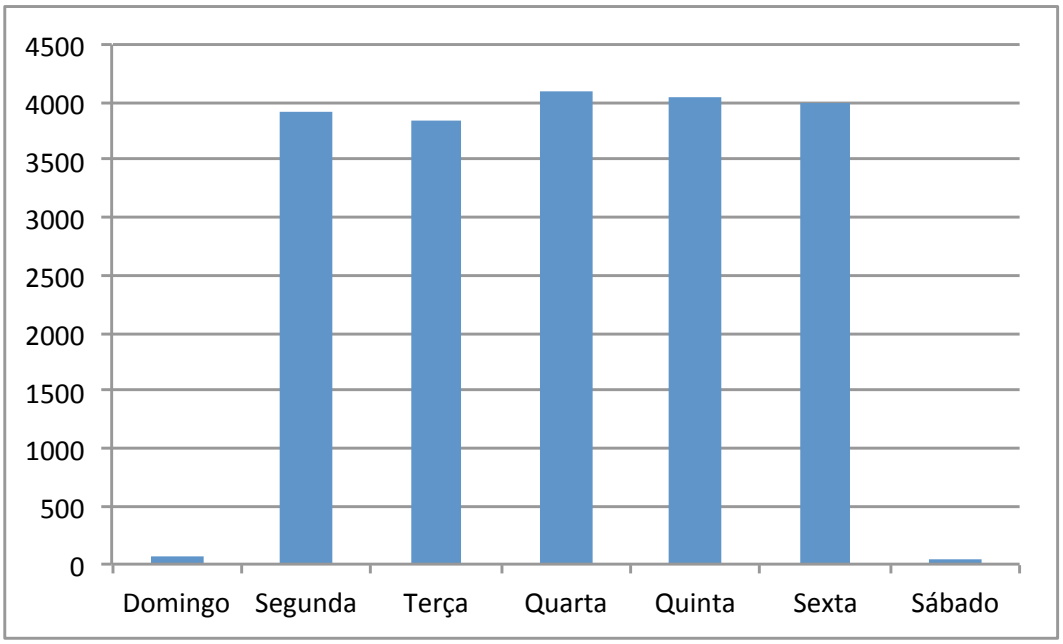


Figura 4.2: Distribuição das Notícias pelos Dias da Semana

O Jornal *Valor Econômico* [23] é um jornal de economia, voltado principalmente para o mercado nacional com notícias redigidas na língua portuguesa. Para a criação do *dataset* utilizado neste trabalho, coletamos automaticamente notícias relacionadas à Petrobras disponíveis no site deste jornal. O site do jornal *Valor Econômico* oferece gratuitamente o serviço de *RSS*, o que facilita a coleta apenas de notícias recentes. Para acesso a base histórica, foi necessário realizar um plano de assinatura que permite acesso ilimitado as notícias do site e as notícias digitalizadas do jornal impresso. Fazendo uso de apenas uma fonte de notícias, evitamos notícias que tratem da mesma informação.

19/11/2013 às 17h54

Petrobras comprova óleo de boa qualidade em poço na área de Franco

Por **Natalia Viri | Valor** Compartilhar:

SÃO PAULO - A Petrobras informou que a conclusão da perfuração de um novo poço na área de Franco comprovou a descoberta de óleo de boa qualidade na região. A área está prevista no contrato de cessão onerosa, no pré-sal da Bacia de Santos.

O poço, informalmente conhecido como Franco Leste, está situado em profundidade de 2 mil metros, a cerca de 200 quilômetros da cidade do Rio de Janeiro e a 7,5 quilômetros a sudeste do poço descobridor. Foi atingida a profundidade final de 5,9 mil metros, após a comprovação de uma coluna de 396 metros de óleo.

"As amostras foram colhidas em reservatórios de espessuras similares às registradas no poço descobridor, comprovando a extensão desses reservatórios com óleo para a região leste do bloco de Franco", disse a Petrobras em comunicado.

Segundo a companhia, a perfuração do poço tem como objetivo delimitar melhor o volume das descobertas em Franco. A fase exploratória tem seu término previsto para setembro de 2014.

Figura 4.3: Notícia sobre a Petrobras

Para realizar a tarefa de coleta das notícias desenvolvemos um *web crawler*, um sistema que possui o objetivo de buscar e extrair documentos para indexação e análise. Esse sistema é capaz de percorrer o ambiente de um site a procura de documentos que possam ser extraídos e processados, a fim de fornecer informação para a criação do *dataset* de notícias. O *crawler* desenvolvido coletou notícias que citassem pelo menos uma vez o nome da empresa Petrobras, ou mencionassem pelo menos uma vez o seu ativo, *PETR4*. Dessa forma, foi realizada uma busca por notícias que atendessem os parâmetros determinados e uma lista de links foi gerada. Para cada link, um

arquivo texto foi criado contendo a informação HTML da página referenciada. Após a visitação de todos os links, foi realizado um pré-processamento nos documentos, removendo a anotação HTML, outros *links*, imagens e tabelas. Desses textos foram extraídas as informações do corpo da notícia, o título e a hora de sua publicação. Na figura 4.3 é apresenta uma notícia que cita o nome da empresa Petrobras.

No total foram coletadas 19.977 notícias, publicadas dentro do período de 01/05/2009 a 30/06/2014. Na figura 4.2 é apresenta a distribuição das notícias pelos dias da semana nas quais foram publicadas. Nota-se uma constância no número de publicações ao longo da semana, com predominância de publicação nos dias úteis. As notícias publicadas nos finais de semana remetem, em sua maior parte, àquelas disponibilizadas por meio impresso e seu conteúdo resume as notícias publicadas nos dias anteriores. Por seu caráter extraordinário e por serem publicadas no final de semana, período de não funcionamento da bolsa de valores, tais notícias não foram utilizadas nessa implementação e criação do *dataset*.

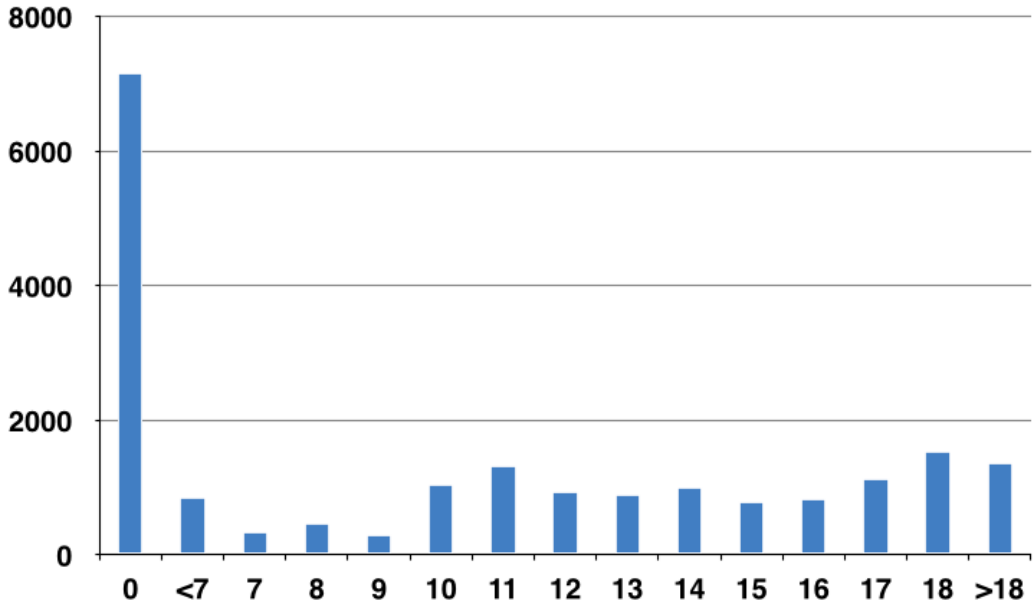


Figura 4.4: Distribuição das Notícias por Hora da Publicação

Na figura 4.4 é apresentada a distribuição de notícias pela hora da publicação. Do total, 7.160 notícias tem como hora de publicação às zero horas. Em sua grande maioria, estas são notícias publicadas na mídia impressa, que são digitalizadas e disponibilizadas no site do jornal para assinantes e usuários. Essas notícias não tem utilidade direta para este trabalho, já que a proposta é analisar o comportamento do mercado financeiro frente a publicação

de uma nova notícia. Dessa forma, para a criação do *dataset*, selecionamos apenas as notícias publicadas durante os dias úteis, com horário de publicação variando entre 7 e 18 horas. Dessa forma, podemos associar diretamente o horário de publicação da notícia com a movimentação da bolsa. No total foram selecionadas 8.969 notícias e sua distribuição pela hora de publicação é apresentada na figura 4.5.

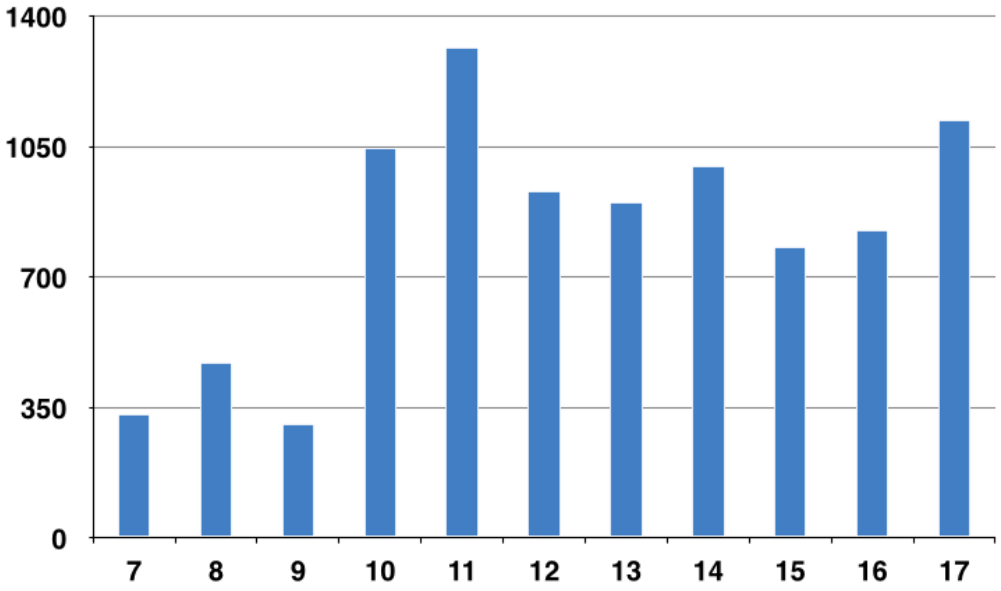


Figura 4.5: Distribuição das Notícias por Hora da Publicação

4.2.2 Dados Históricos

A BMF&Bovespa é uma companhia que administra mercados organizados de títulos e valores imobiliários, além de prestar serviços de registro, compensação e liquidação, atuando, principalmente, como contraparte central garantidora da liquidação financeira das operações realizadas em seus ambientes[24]. Sua origem remete ao surgimento da Bolsa Livre, em 1891, ainda no governo de Marechal Deodoro da Fonseca, se consolidando em 1967, já com o nome de Bovespa - Bolsa de Valores de São Paulo. Em 2009, após uma grande modernização, passa a realizar todas as negociações apenas por meio eletrônio, extinguindo os antigos pregões realizados por viva voz.

O mercado de ações é o principal meio de de captação de recursos de empresas de capital aberto. Através da bolsa de valores, grandes empresas tem a oportunidade de arrecadar fundos para realizar projetos, com a venda de ações da empresa no mercado aberto. No contexto brasileiro, a Bovespa assume

o papel de organizar este mercado, no intuito de oferecer um mecanismo seguro para a realização de compra e venda de papéis de empresas do cenário nacional. Dentre estas grandes empresas de capital aberto, se destaca a Petrobras, companhia petrolífera brasileira, que participa da bolsa de valores com o ativo *PETR4*.

Neste trabalho foi criado uma base de dados de informações *intraday* do ativo *PETR4* na BMF&Bovespa, dentro do período de 12/2006 à 06/2014. Esta base de informações numéricas sobre o mercado se faz necessária para a realização da tarefa de anotação. A Tabela 4.1 abaixo ilustra uma série histórica de quinze valores.

Tabela 4.1: Informação Numérica *Intraday*

Horário	*Abertura*	*Máx*	*Mín*	*Fecham*	*Quant*
2013-10-10 10:05:00	18,26	18,3	18,25	18,3	2071070
2013-10-10 10:06:00	18,3	18,31	18,27	18,28	695191
2013-10-10 10:07:00	18,29	18,29	18,26	18,27	358110
2013-10-10 10:08:00	18,28	18,3	18,26	18,29	942875
2013-10-10 10:09:00	18,29	18,29	18,28	18,28	67662
2013-10-10 10:10:00	18,28	18,28	18,27	18,28	286958
2013-10-10 10:11:00	18,28	18,28	18,28	18,28	100540
2013-10-10 10:12:00	18,28	18,29	18,28	18,29	409537
2013-10-10 10:13:00	18,29	18,29	18,28	18,29	157292
2013-10-10 10:14:00	18,28	18,28	18,26	18,28	487818
2013-10-10 10:15:00	18,28	18,3	18,27	18,29	371295
2013-10-10 10:16:00	18,3	18,3	18,29	18,29	440815
2013-10-10 10:17:00	18,29	18,3	18,28	18,28	373042
2013-10-10 10:18:00	18,28	18,29	18,28	18,29	166405
2013-10-10 10:19:00	18,28	18,28	18,24	18,24	887138

Cada linha da tabela apresenta os valores negociados a cada minuto, dentro de um dia de negociação. Dentre as diversas informações disponibilizadas, são extraídas as informações da série referente a cotação do preço de abertura (coluna Abertura), preço máximo (Máx), preço mínimo (Mín), preço de fechamento (Fecham) e quantidade negociada (Quant) para cada minuto. Dessa forma, o preço de abertura é o valor negociado no exato início do minuto e o valor de fechamento o preço no instante final. Os preços máximo e mínimo correspondem ao maior e menor valor respectivamente dentro do minuto de negociação, incluindo os valores de abertura e fechamento. A coluna de quantidade apresenta as negociações realizadas dentro deste período. Ao se verificar um aumento demasiado da quantidade negociada, não é possível inferir que há

necessariamente uma grande agitação no mercado, pois este movimento pode ser consequência da compra ou venda de um grande investidor.

4.3 Representação da Notícia

O sucesso de qualquer classificador na geração de resultados precisos dentro da tarefa de classificação de notícias depende da forma em que os textos utilizados são representados. Para que os algoritmos de classificação possam determinar uma função aprendizado, é necessário prover diversas informações sobre as notícias.

Neste trabalho utilizamos atributos estruturais e linguísticos dentro da representação por presença dos textos. Dado essa forma de representação, foi criado uma lista associando cada atributo a um índice único. Caso o atributo esteja presente no texto analisado, será atribuído ao vetor de representação, no índice referente ao atributo, o valor 1 e caso ausente o valor 0.

Dentre os atributos gerados, podemos citar a geração de n-gramas de palavras, anotação morfossintática e classificação de *chunks*.

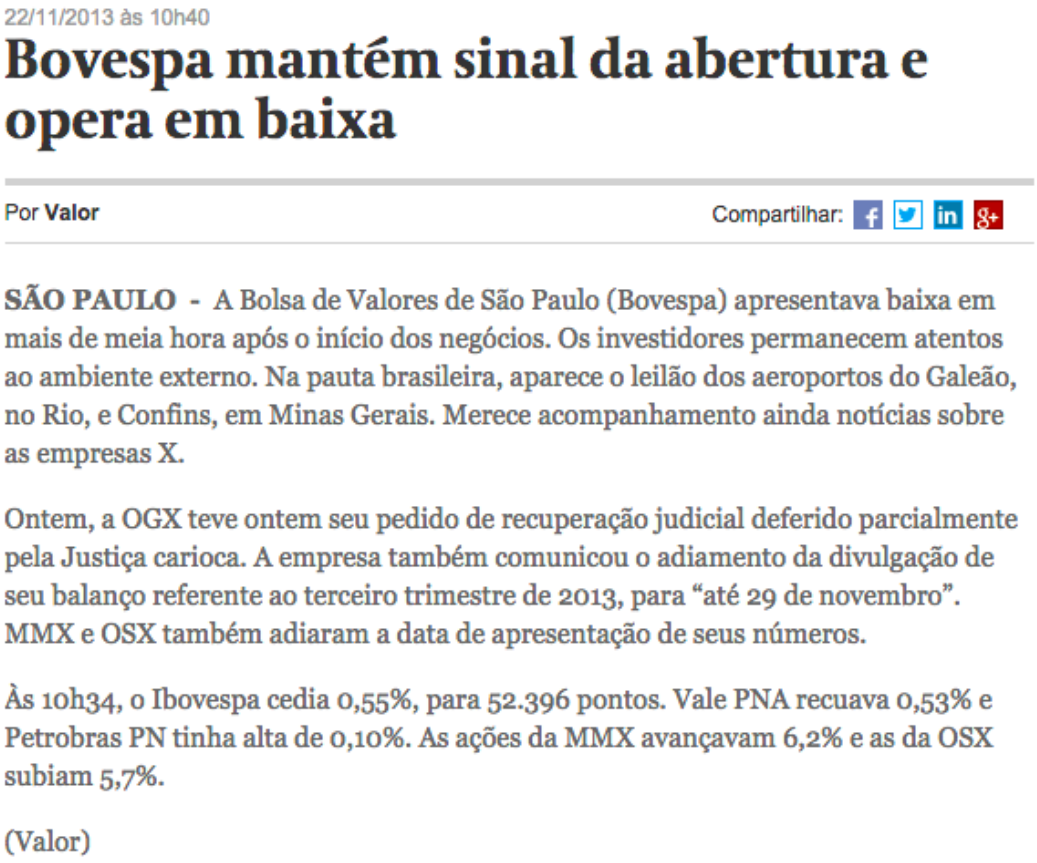
Para a geração dos atributos linguísticos, anotação morfossintática e *chunks*, foi utilizado o FEXT [25], sistema desenvolvido pelo laboratório LEARN da PUC-Rio. O FEXT utiliza o algoritmo de aprendizado ETL (*Entropy Guided Transformation Learning*), que utiliza a estratégia de combinar as vantagens da árvore de decisão com TBL (*Transformation Based Learning*).

A seguir, discutiremos os métodos de extração de atributos utilizados neste trabalho.

4.3.1 Seleção de frases

Muitas das notícias coletadas para esse trabalho apresentam informações sobre várias empresas em uma mesma notícia. A figura abaixo 4.6 ilustra uma das notícias utilizadas.

Nesse exemplo, o trecho da notícia fala sobre diversas empresas, algumas com valorização e outras com desvalorização. Esse tipo de informação pode causar ruído e piorar a acurácia do classificador. Nesse sentido, foi desenvolvido um seletor de frases que filtra das notícias somente frases relativas a uma determinada empresa. Neste seletor, por exemplo, são utilizadas as seguintes palavras chaves: *Petrobras*, o nome da empresa, *PETR4*, o ativo da empresa e *petróleo*, produto produzido pela empresa. Note que os termos utilizados não foram escolhidos por um especialista, apenas representam informações gerais



22/11/2013 às 10h40

Bovespa mantém sinal da abertura e opera em baixa

Por **Valor** Compartilhar:

SÃO PAULO - A Bolsa de Valores de São Paulo (Bovespa) apresentava baixa em mais de meia hora após o início dos negócios. Os investidores permanecem atentos ao ambiente externo. Na pauta brasileira, aparece o leilão dos aeroportos do Galeão, no Rio, e Confins, em Minas Gerais. Merece acompanhamento ainda notícias sobre as empresas X.

Ontem, a OGX teve ontem seu pedido de recuperação judicial deferido parcialmente pela Justiça carioca. A empresa também comunicou o adiamento da divulgação de seu balanço referente ao terceiro trimestre de 2013, para "até 29 de novembro". MMX e OSX também adiaram a data de apresentação de seus números.

Às 10h34, o Ibovespa cedia 0,55%, para 52.396 pontos. Vale PNA recuava 0,53% e Petrobras PN tinha alta de 0,10%. As ações da MMX avançavam 6,2% e as da OSX subiam 5,7%.

(Valor)

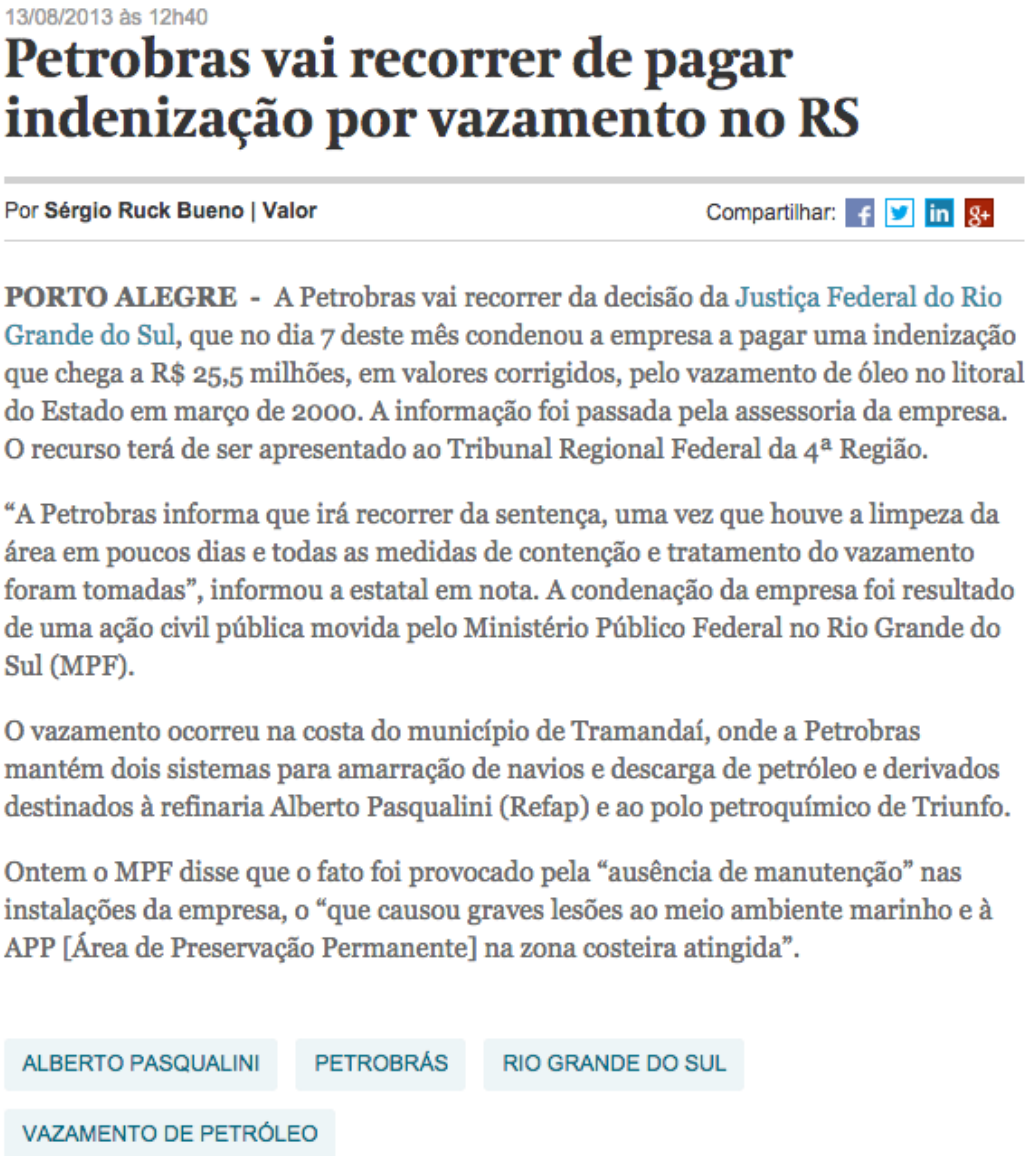
Figura 4.6: Seleção de Frases

da empresa. Dessa forma, o trabalho aqui apresentado continua genérico para ser utilizado com outros ativos.

4.3.2 Atributos Estruturais

Os atributos estruturais trazem informações sobre a estrutura de um texto. Determinados atributos podem indicar o grau de relevância de determinada notícia e quantidade de informação disponível sobre determinada empresa.

Neste trabalho utilizamos três categorias de atributos estruturais: número médio de sentenças por notícia, presença de referência a empresa no título e nas *tags* da notícia e a frequência de referências a empresa e ao ativo no corpo da notícia. Na figura 4.7 é apresentado um exemplo de notícia com referências a empresa no título, no corpo da notícia e na *tag*.



13/08/2013 às 12h40

Petrobras vai recorrer de pagar indenização por vazamento no RS

Por **Sérgio Ruck Bueno | Valor** Compartilhar:

PORTO ALEGRE - A Petrobras vai recorrer da decisão da Justiça Federal do Rio Grande do Sul, que no dia 7 deste mês condenou a empresa a pagar uma indenização que chega a R$ 25,5 milhões, em valores corrigidos, pelo vazamento de óleo no litoral do Estado em março de 2000. A informação foi passada pela assessoria da empresa. O recurso terá de ser apresentado ao Tribunal Regional Federal da 4ª Região.

"A Petrobras informa que irá recorrer da sentença, uma vez que houve a limpeza da área em poucos dias e todas as medidas de contenção e tratamento do vazamento foram tomadas", informou a estatal em nota. A condenação da empresa foi resultado de uma ação civil pública movida pelo Ministério Público Federal no Rio Grande do Sul (MPF).

O vazamento ocorreu na costa do município de Tramandaí, onde a Petrobras mantém dois sistemas para amarração de navios e descarga de petróleo e derivados destinados à refinaria Alberto Pasqualini (Refap) e ao polo petroquímico de Triunfo.

Ontem o MPF disse que o fato foi provocado pela "ausência de manutenção" nas instalações da empresa, o "que causou graves lesões ao meio ambiente marinho e à APP [Área de Preservação Permanente] na zona costeira atingida".

ALBERTO PASQUALINI PETROBRÁS RIO GRANDE DO SUL

VAZAMENTO DE PETRÓLEO

Figura 4.7: Atributos Estruturais

4.3.3 Bag of Words

Uma forma de representar uma notícia é utilizar apenas as palavas nela contida. Este é o método de extração mais simples e conhecido como o saco de palavras (*bag of words*). Neste método cada palavra do texto é utilizada como um atributo, ignorando os sinais de pontuação.

A ocorrência de um termo em um documento estabelece uma relação dele com os demais. As relações entre as palavras e os documentos podem ser apresentadas pela presença de um termo no texto. Na forma matricial, teremos como exemplo a matriz 4-1, onde cada elemento $p_{i,j}$ representa a presença do

termo k_i no documento d_j. Caso o valor de $p_{i,j}$ seja 1, o termo k_i existe no documento d_j, caso contrário será 0.

$$\begin{array}{c} \\ \mathrm{k}_1 \\ \mathrm{k}_2 \\ \vdots \\ \mathrm{k}_m \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{d}_1 & \mathrm{d}_2 & \cdots & \mathrm{d}_n \end{array} \\ \begin{pmatrix} \mathrm{p}_{1,1} & \mathrm{p}_{1,2} & \cdots & \mathrm{p}_{1,n} \\ \mathrm{p}_{2,1} & \mathrm{p}_{2,2} & \cdots & \mathrm{p}_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{p}_{m,1} & \mathrm{p}_{m,2} & \cdots & \mathrm{p}_{m,n} \end{pmatrix} \end{array} \tag{4-1}$$

4.3.4 Atributos N-Gramas

Um n-grama é uma combinação de n termos. Estes termos podem ser agrupados dois-a-dois, um bigrama; em três-a-três, um trigrama ou n-grama para qualquer outro tamanho. Considere o exemplo abaixo:

"Petrobras negocia o valor da cessão."

Dessa forma, para este exemplo, podemos gerar os seguintes bigramas: *"Petrobras negocia"*, *"negocia o"*, *"o valor"*, *"valor da"*, *"da cessão"*. E podemos gerar os seguintes trigramas: *"Petrobras negocia o"*, *"negocia o valor"*, *"o valor da"*, *"valor da cessão"*.

Esse tipo de representação apresenta várias vantagens frente a outros modelos, agregando mais informação a um atributo. Considere o seguinte trecho extraído de uma notícia:

"Ações da Petrobras sobem 12%."

Neste exemplo, seria gerado o bigrama *"Petrobras sobe"*, um atributo informativo sobre a valorização do ativo e ainda o trigrama *"Petrobras sobe 12"*, atributo com informação quantitativa do valor da valorização.

4.3.5 Anotação Morfossintática

A tarefa de anotação morfossintática consiste no processo de anotação das palavras de um texto com uma etiqueta que denota sua classe gramatical. Esta etiqueta é conhecida também como etiqueta POS, da expressão em inglês Part-of-Speech. A tarefa classifica as palavras em categorias baseado nas regras

do contexto em que essas aparecem [17]. Na Tabela 4.2, é apresentada uma sentença com sua anotação de POS. Cada linha corresponde a uma palavra e a coluna POS ao atributo gerado.

Tabela 4.2: POS Tag

Palavra	*POS*
Justiça	NPROP
abre	V
processo	N
contra	PREP
ex-diretor	N
de	PREP
a	ART
Petrobras	NPROP
.	.

Para estre trabalho, consideramos todas as classes morfossintáticas descritas no apêndice A.

4.3.6 Chunk

Text chunking, ou segmentação textual, é o processo de identificar e classificar sequências disjuntas de uma sentença, formadas por palavras sintaticamente relacionadas. Esse conjunto tipicamente constitui sintagmas nominais, sintagmas verbais, sintagmas adjetivais e sintagmas preposicionais [17]. Na Tabela 4.3 é apresentada uma sentença com sua anotação *chunk*.

Tabela 4.3: Chunk

Palavra	*CHUNK*
Justiça	B-NP
abre	B-VP
processo	B-NP
contra	B-PP
ex-diretor	B-NP
de	B-PP
a	B-NP
Petrobras	I-NP
.	B-NP

4.4 Anotação Automática

Alguns métodos de predição do comportamento de ações são baseados apenas no histórico dos preços, utilizando-se da análise técnica [3]. Este trabalho, por sua vez, combina os valores de preços das ações com o conteúdo das notícias. Na etapa da anotação automática utilizamos uma heurística para encontrar uma relação entre os preços e determinada notícia.

Um aspecto importante na tarefa de anotação é o horizonte de predição, que diz respeito à janela de tempo após a liberação da notícia. Para a predição a curto prazo, o intervalo de tempo pode ser de cinco a quinze minutos [8]. Dessa forma, as informações do mercado dentro desse intervalo de tempo são utilizadas para anotação.

Outro aspecto importante a ser considerado é o número de classes a serem utilizadas. Uma classe indica a relação entre a notícia atual e o comportamento futuro de uma ação. O número total de classes escolhidas determina critérios diferentes para a anotação de uma nova notícia [26]. Para essa tarefa, conforme apresentado, definimos três classes: “alta”, “baixa” e “neutra”. Uma notícia é considerada ”alta”se após a sua liberação os preços do ativo associado sobem e, analogamente ”baixa”se após a liberação os preços caem. No caso de nenhuma ou baixa variação, a notícia é considerada ”neutra”.

Para cada notícia, é computada a maior e a menor variação dos preços médios de minuto a minuto em uma janela de quinze minutos. Com essa informação, se sua variação for positiva e maior que 0,30%, então a notícia é anotada com a classe ”alta”. Caso a variação seja negativa e maior que 0,30% então é anotada com a categoria ”baixa”. Nos demais casos, se nenhum dos critérios anteriores for preenchido, a notícia é anotada com a categoria ”neutra”.

4.5 Seleção de Atributos

Dado a alta dimensão do espaço de atributos, se faz necessário o uso de um seletor para selecionar um subconjunto de atributos mais informativos, reduzindo a dimensionalidade do problema e assim melhorando a precisão do classificador. Neste trabalho, utilizamos uma implementação do perceptron esparso para a seleção de atributos desenvolvida por Motta [27]. A seguir, no algoritmo 4.1, é apresentamos o pseudo-código implementado.

Esta variação do algoritmo original presume que somente atributos ditos relevantes devem permanecer no modelo final. Inicialmente, todos os atributos são considerados irrelevantes. Dessa forma, para ser considerado relevante, é

necessário que o atributo participe em pelo menos L atualizações na matriz de pesos [28].

Algoritmo 4.1: Perceptron Multiclasse Esparso

Entrada:
- $\mathcal{D}$ — Dataset de treino
- $\mathbf{w}_0$ — Pesos iniciais
- U_0 — Contagens iniciais
- $ÉPOCAS$ — Número de épocas
- L — Limiar de seleção
- $Reset\,Contagens$ — Flag que indica se U deve ser zerado a cada época

Saída: $\mathbf{K} = \{k \mid u_k \geq L\}$ Atributos selecionados

1: $\mathbf{w} \leftarrow \mathbf{w}_0$
2: $\mathbf{U} \leftarrow \mathbf{U}_0$
3: $t \leftarrow 0$
4: **while** $t < ÉPOCAS$ **do**
5: **for each** $(x, y) \in \mathcal{D}$ **do**
6: $\hat{y} \leftarrow arg\,max_{1 \leq k \leq C}\{w_k[\mathbf{U} \geq L].\mathbf{x}\}$
7: **if** $\hat{y} \neq y$ **then**
8: $\mathbf{w}_y \leftarrow \mathbf{w}_y + \mathbf{x}$
9: $\mathbf{w}_{\hat{y}} \leftarrow \mathbf{w}_{\hat{y}} - \mathbf{x}$
10: **for each** i **s.t.** $x_{ik} \neq 0$ **do**
11: $U_i \leftarrow U_i + 1$
12: **end for**
13: **end if**
14: **end for**
15: **if** $Reset\,Contagens$ **then**
16: **for each** i **s.t.** $U_i < L$ **do**
17: $U_i \leftarrow 0$
18: **end for**
19: **end if**
20: $t \leftarrow t + 1$
21: **end while**
22: $\mathbf{K} \leftarrow \{k \mid U_k \geq L\}$
23: **return** $\mathbf{K}$

4.6 Considerações Finais

Neste capítulo, apresentamos a tarefa de predição de comportamento do mercado financeiro, a criação dos *datasets* utilizados, a engenharia de atributos, a tarefa de anotação automática e a seleção de atributos. Na primeira seção, ficou claro que esta tarefa pode ser transformada em um problema de classificação de textos utilizando um algoritmo de Aprendizado de Máquina Supervisionado. Nas seções seguintes, apresentamos o processo de coleta de dados e de representação da notícia. Além dos atributos amplamente utilizados

pela literatura, geramos novos atributos linguísticos. Na seção Anotação Automática, apresentamos a técnica utilizada para anotar as notícias, adaptada ao nosso problema. Por fim, na última seção, apresentamos a técnica utilizada para selecionar atributos, reduzindo a dimensionalidade do problema.

5
Experimentos

Nesse capítulo apresentamos os experimentos realizados e a seguir detalhamos o impacto de cada um desses. Como os conjuntos de notícias utilizados não possuem a mesma quantidade de dados por classe, é necessário manter a proporção tanto nos conjuntos de treino como nos conjuntos de teste. Para todos os experimentos realizados, verificamos a acurácia utilizando validação cruzada 10-*fold*. Dessa forma, separamos 10 conjuntos, cada um contendo todas as notícias, sendo 90% utilizadas na etapa de treinamento e 10% na etapa de teste.

5.1
Impacto do uso de atributos estruturais

Tabela 5.1: Impacto do uso de atributos estruturais

Modelo	*Acurácia (%)*
Atributos Estruturais	55,49
Baseline	50,84

Atributos estruturais não são intuitivamente informativos, mas podem revelar padrões dos textos e melhorar a performance do classificador. Realizamos testes com as três categorias de atributos estruturais propostas: número médio de sentenças, presença de referência a empresa no título e nas *tags* das notícias e frequência de referências a empresa e ao ativo no corpo da notícia. Na tabela 5.1, verificamos uma ligeira melhora em relação ao modelo *baseline*.

O resultado apresentado demonstra que apenas atributos estruturais não são suficientes para a representação de um documento, sendo necessário a combinação com outros atributos.

5.2
Impacto do uso de PLN

Para este experimento, foram acrescentados os atributos linguísticos de anotação morfossintáticas e *chunk*. Os novos atributos são combinados

Tabela 5.2: Impacto do uso de PLN

Modelo	*Acurácia (%)*
1,2,3-gramas de palavras	65,9
Palavra + POS + Chunk	66,53

com as palavras que representam, gerando os pares palavra_EtiquetaMorfo e palavra_EtiquetaChunk, além do trio palavra_EtiquetaMorfo_EtiquetaChunk.

Apesar de gerarem um número quatro vezes maior de atributos, nossos experimentos indicam uma melhora na acurácia, demonstrando que os atributos são informativos. Na tabela 5.2 pode-se observar o ganho com a inclusão dos novos atributos.

5.3 Impacto do uso de n-gramas de etiquetas morfossintática

Tabela 5.3: Impacto do uso de n-gramas de etiquetas morfossintáticas

Modelo	*Acurácia (%)*
N-gramas POS	66,60
Palavra + POS + Chunk	66,53

Nesse experimento, combinamos os atributos morfossintáticos em bigramas e trigramas. São selecionados os atributos das palavras na ordem em que aparecem no texto. Dessa forma, criamos os pares EtiquetaMorfo_EtiquetaMorfo e os trios EtiquetaMorfo_EtiquetaMorfo_EtiquetaMorfo. Pode-se observar na tabela 5.3 que a adição desses atributos não contribui de forma significativa, produzindo mais ruído.

5.4 Impacto do uso do seletor de atributos

Tabela 5.4: Impacto do uso de seletor de atributos

Modelo	*Acurácia (%)*
Perceptron Esparso	68,57
N-gramas POS	66,60
Palavra + POS + Chunk	66,53

Utilizamos uma implementação [1] do perceptron esparso para a seleção

[1] Comunicação pessoal de Eduardo Motta, em 16 de agosto de 2014, recebida por correio eletrônico.

de atributos. Os atributos binários são processados pelo perceptron esparso e somente alguns atributos mais informativos são selecionados. Dessa forma, há a diminuição da cardinalidade dos atributos originais.

Para avaliar a eficiência do seletor de atributos, executamos o experimento utilizando o modelo que contempla os atributos estruturais, de n-grama das palavras, os atributos de PLN e de n-gramas das etiquetas morfossintáticas. Na tabela 5.4 é apresentado o melhor resultado variando o parâmetro de *threshold*, obtendo uma redução do número de atributos em 55%. O uso do seletor de atributos resulta em uma diminuição de erro de 5,89% se compararmos com o modelo sem seleção.

5.5 Resultados

Tabela 5.5: Melhor resultado

Modelo	*Precision (%)*	*Recall (%)*	$F_{\beta=1}(\%)$
1,2,3-gramas + POS + Chunk	0,69	0,69	0,66

Na tabela 5.5, apresentamos o melhor resultado, fruto da combinação de vários atributos. O melhor modelo é formado por atributos estruturais, 1,2,3-gramas de palavras, bigramas e trigramas da combinação palavra_etiqueta com os atributos PLN.

Tabela 5.6: Resultados por Classe

	Precision (%)	*Recall (%)*	$F_{\beta=1}(\%)$
Neutra	0,76	0,39	0,52
Alta	0,65	0,87	0,74
Baixa	0,68	0,90	0,78

A tabela 5.6 apresenta os resultados distribuídos por cada classe. Embora a classe Neutra seja a mais frequente, seu resultado frente as outras classes foi pior. Uma possível explicação para esse resultado se deve ao fato de que grande número de notícias dessa classe é caracterizada por ser relatório técnico, não agregando informação ao classificador.

Por fim, os resultados mostram a importância da utilização de outras métricas para a avaliação dos resultados. Na maioria dos trabalhos relacionados, a única métrica utilizada é a acurácia. Por se tratar de um conjunto desbalanceado de dados, é importante demonstrar que o classificador está conseguindo aprender com o conjunto de treino.

6
Conclusões

Neste capítulo apresentamos nossas considerações finais. Na primeira seção, descrevemos o problema abordado nessa dissertação, predição do comportamento do mercado financeiro utilizando notícias escritas em português. Na seção seguinte, descrevemos nossas principais contribuições. Por fim, na seção Trabalhos Futuros, dissertamos sobre melhorias e sugestões para o trabalho abordado.

6.1
Visão Geral

Nesta dissertação investigamos como traduzir o conteúdo de notícias sobre o mercado financeiro em predição do comportamento da bolsa de valores. Nosso objetivo é classificar as notícias jornalísticas a fim de encontrar relação com o mercado e dessa forma predizer seu comportamento nas posições de alta, baixa e neutra.

A metodologia padrão consiste em classificar as notícias sobre o mercado. Uma classe indica uma relação estrita entre a publicação de uma notícia e o comportamento de uma ação frente a essa publicação. No processo de anotação das classes, utiliza-se uma heurística para achar uma relação entre a variação do preço de um ativo e a notícia corrente. Após a etapa de anotação, treina-se um classificador para aprender tais relações. Dada a publicação de uma nova notícia, a saída do classificador será a predicação do comportamento do mercado frente a essa nova publicação. Ao contrário das estratégias estudadas, utilizamos também atributos linguísticos de anotação morfossintática e *chunk*, combinados com atributos estruturais relacionados ao tamanho da notícia e a presença de termos chaves.

6.2
Contribuições

A abordagem padrão, utilizada nos trabalhos estudados, utiliza apenas atributos estruturais para a representação das notícias. Em vista disto, propomos utilizar atributos linguísticos combinados com atributos estruturais, no

intuito de fornecer mais informações ao classificador. A aplicação e uso de atributos de anotação morfossintática e *chunk* produziram um ganho em relação à representação *baseline*. O uso de atributos linguísticos demonstrou ser mais representativo, obtendo dessa forma melhores resultados.

6.3 Trabalhos Futuros

Como trabalhos futuros, podemos avaliar o uso de outros algoritmos de Aprendizado de Máquina para a tarefa de classificação com o intuito de melhorar a predição.

Adicionalmente, pode-se desenvolver uma estratégia de múltiplos classificadores. Cada classificador trabalharia em uma determinada fração de tempo durante um dia de negociação, sendo modelado com notícias publicadas apenas dentro deste intervalo. Dessa forma, os classificadores seriam específicos para determinados horários e mais sensíveis a variações.

Uma outra direção interessante, seria buscar ativos correlacionados à Petrobras, dentro de uma mesma notícia ou mesmo em notícias diferentes. Através dessa associação, verificar a analisar o impacto no preço do ativo PETR4.

Por fim, investigar o uso de novas formas de representação do texto, tais como *Semantic Role Labeling* e *Named Entities*.

Referências Bibliográficas

[1] JOACHIMS, T. Text categorization with support vector machines: Learning with many relevant features. In: NDELLEC, C.; ROUVEIROL, C. (Ed.). **Machine Learning: ECML-98**. Springer Berlin Heidelberg, 1998, (Lecture Notes in Computer Science, v. 1398). p. 137–142. ISBN 978-3-540-64417-0. Disponível em: <http://dx.doi.org/10.1007/BFb0026683>.

[2] MALKIEL, B. **A Random Walk Down Wall Street: The Time-Tested Strategy for Successful Investing (Tenth Edition)**. W. W. Norton, 2011. ISBN 9780393081695. Disponível em: <http://books.google.com.br/books?id=O8x1YpBp6WYC>.

[3] FORTUNY, E. J. de et al. Evaluating and understanding text-based stock price prediction models. **Information Processing & Management**, v. 50, n. 2, p. 426 – 441, 2014. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457313001143>.

[4] WUTHRICH, B. et al. Daily stock market forecast from textual web data. In: **Systems, Man, and Cybernetics, 1998. 1998 IEEE International Conference on**. [S.l.: s.n.], 1998. v. 3, p. 2720–2725 vol.3. ISSN 1062-922X.

[5] LAVRENKO, V. et al. Language models for financial news recommendation. In: **Proceedings of the Ninth International Conference on Information and Knowledge Management**. New York, NY, USA: ACM, 2000. (CIKM '00), p. 389–396. ISBN 1-58113-320-0. Disponível em: <http://doi.acm.org/10.1145/354756.354845>.

[6] GIDOFALVI, G. **Using News Articles to Predict Stock Price Movements**. [S.l.], 2001.

[7] PERAMUNETILLEKE, D.; WONG, R. K. Currency exchange rate forecasting from news headlines. In: **Proceedings of the 13th Australasian Database Conference - Volume 5**. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2002. (ADC '02), p. 131–139. ISBN 0-909925-83-6. Disponível em: <http://dl.acm.org/citation.cfm?id=563906.563921>.

[8] MITTERMAYER, M.-A.; KNOLMAYER, G. Newscats: A news categorization and trading system. In: **Data Mining, 2006. ICDM '06. Sixth International Conference on**. [S.l.: s.n.], 2006. p. 1002–1007. ISSN 1550-4786.

[9] LI, X. et al. Improving stock market prediction by integrating both market news and stock prices. In: HAMEURLAIN, A. et al. (Ed.). **Database and Expert Systems Applications**. Springer Berlin Heidelberg, 2011, (Lecture Notes in Computer Science, v. 6861). p. 279–293. ISBN 978-3-642-23090-5. Disponível em: <http://dx.doi.org/10.1007/978-3-642-23091-2$_24$¿.

[10] BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval: The Concepts and Technology Behind Search**. Addison Wesley, 2011. ISBN 9780321416919. Disponível em: <http://books.google.com.br/books?id=HbyAAAAACAAJ>.

[11] MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. **Introduction to Information Retrieval**. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715. Disponível em: <http://www-nlp.stanford.edu/IR-book/>.

[12] MITCHELL, T. M. **Machine Learning**. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

[13] ABBASI, A.; CHEN, H.; SALEM, A. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. **ACM Trans. Inf. Syst.**, ACM, New York, NY, USA, v. 26, n. 3, p. 12:1–12:34, jun. 2008. ISSN 1046-8188. Disponível em: <http://doi.acm.org/10.1145/1361684.1361685>.

[14] MANNING, C. D.; SCHüTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge, MA, USA: MIT Press, 1999. ISBN 0-262-13360-1.

[15] TSOUMAKAS, G.; KATAKIS, I. Multi-label classification: An overview. **Int J Data Warehousing and Mining**, v. 2007, p. 1–13, 2007.

[16] RUSSELL, S.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 3rd. ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2009. ISBN 0136042597, 9780136042594.

[17] JURAFSKY, D.; MARTIN, J. **Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition**. Pearson Prentice Hall, 2009. (Prentice

Hall series in artificial intelligence). ISBN 9780131873216. Disponível em: <http://www.cs.colorado.edu/ martin/slp.html>.

[18] PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

[19] BOSER, B. E.; GUYON, I. M.; VAPNIK, V. N. A training algorithm for optimal margin classifiers. In: **Proceedings of the Fifth Annual Workshop on Computational Learning Theory**. New York, NY, USA: ACM, 1992. (COLT '92), p. 144–152. ISBN 0-89791-497-X. Disponível em: <http://doi.acm.org/10.1145/130385.130401>.

[20] FORMAN, G. An extensive empirical study of feature selection metrics for text classification. **J. Mach. Learn. Res.**, JMLR.org, v. 3, p. 1289–1305, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944974>.

[21] YONG-FENG, S.; YAN-PING, Z. Comparison of text categorization algorithms. **Wuhan University Journal of Natural Sciences**, Wuhan University, v. 9, n. 5, p. 798–804, 2004. ISSN 1007-1202. Disponível em: <http://dx.doi.org/10.1007/BF02831684>.

[22] FAN, R.-E. et al. Liblinear: A library for large linear classification. **J. Mach. Learn. Res.**, JMLR.org, v. 9, p. 1871–1874, jun. 2008. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1390681.1442794>.

[23] VALOR. **Jornal Valor Econômico**. Acesso em 29 Jul 2014. Disponível em: <http://www.valor.com.br/>.

[24] BOVESPA. **BMFBOVESPA**. Acesso em 29 Jul 2014. Disponível em: <http://http://www.bmfbovespa.com.br/>.

[25] FERNANDES E.; MILIDI, R. D. S. C. Portuguese language processing service. In: **Web in Ibero-America Alternate Track of the 18th International World Wide Web Conference**. [S.l.: s.n.], 2009.

[26] FUNG, G.; YU, J.; LAM, W. News sensitive stock trend prediction. In: CHEN, M.-S.; YU, P.; LIU, B. (Ed.). **Advances in Knowledge Discovery and Data Mining**. [S.l.]: Springer Berlin Heidelberg, 2002, (Lecture Notes in Computer Science, v. 2336). p. 481–493. ISBN 978-3-540-43704-8.

[27] MOTTA, E. N. Indução e seleção incrementais de atributos no aprendizado supervisionado. Tese(Doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, 2014.

[28] ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological Review**, v. 65, n. 6, p. 386–408, nov. 1958.

A
Conjunto de Etiquetas Morfossintáticas

Tabela A.1: Etiquetas Morfossintáticas

Classe Morfossintática	*Etiqueta*
Adjetivo	ADJ
Advérbio	ADV
Artigo (definido ou indefinido)	ART
Símbolo de moeda corrente	CUR
Conjunção coordenativa	KC
Conjunção subordinativa	KS
Nome	N
Nome Próprio	NPROP
Numeral	NUM
Particípio	PCP
Palavra denotativa	PDEN
Preposição	PREP
Pronome Adjetivo	PROPADJ
Pronome pessoal	PROPESS
Pronome relativo conectivo subordinativo	PRO-KS-REL
Pronome substantivo	PROSUB
Verbo	V
Verbo Auxiliar	VAUX

B
Conjunto de Etiquetas Chunk

Tabela B.1: Etiquetas Chunk

Classe Morfossintática	*Etiqueta*
Início de chunk nominal	B-NP
Início de chunk preposicional	B-PP
Início de chunk verbal	B-VP
Contido em um chunk nominal	I-NP
Contido em um chunk preposicional	I-PP
Contido em um chunk verbal	I-VP
Fora de qualquer chunk	O