



UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
ESCOLA DE INFORMÁTICA APLICADA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Mining StockTec: Predição de preço de ações através de mineração de dados e análise de sentimentos.

Nome dos autores:

Gustavo Mendonça do Rio Branco

Marcos André Rosendo Barroso

Nome das Orientadoras:

Kate Revoredo

Fernanda Baião

Maio/2014

Mining StockTec: Predição de preço de ações através de mineração de dados e análise de sentimentos.

Projeto de Graduação apresentado à Escola de Informática Aplicada da Universidade Federal do Estado do Rio de Janeiro (UNIRIO) para obtenção do título de Bacharel em Sistemas de Informação.

Nome dos autores:

Gustavo Mendonça do Rio Branco
Marcos André Rosendo Barroso

Nome das Orientadoras:

Kate Revoredo e Fernanda Baião

Mining StockTec: Predição de preço de ações através de mineração de dados e análise de sentimentos.

Aprovado em ____/_____/____

BANCA EXAMINADORA

Prof. Fernanda Baião, D.Sc. (UNIRIO)

Prof. Kate Revoredo, D.Sc. (UNIRIO)

Prof. Marcio Barros, D.Sc. (UNIRIO)

Prof. Sean Siqueira, D.Sc. (UNIRIO)

Os autores deste Projeto autorizam a ESCOLA DE INFORMÁTICA APLICADA da UNIRIO a divulgá-lo, no todo ou em parte, resguardando os direitos autorais conforme legislação vigente.

Rio de Janeiro, ____ de _____ de _____

Marcos André Rosendo Barroso

Gustavo Mendonça do Rio Branco

AGRADECIMENTOS

Marcos Barroso:

Não seria possível chegar até este momento sem a ajuda plena de meus pais Mario e Mariza, é deles que advêm as forças necessárias para continuar estudando e não desistir de galgar espaços maiores perante a sociedade. Logo, inicio agradecendo aos meus pais por sempre acreditar e me apoiar em todas as decisões.

Agradeço também à minha irmã Liliane, por ser um exemplo de conquista, inteligência e perseverança em minha vida demonstrando que é possível chegar a qualquer lugar através do esforço e dedicação.

À minha namorada Carla, que com seu amor soube me acalmar nos momentos de nervosismo e sempre encorajar-me em cada atividade exercida. Aos meus amigos, que sempre creditaram a mim confiança para testar em seus computadores os mais diversos experimentos.

Agradeço à Universidade Federal do Estado do Rio de Janeiro (UNIRIO) e a Escola de Informática Aplicada (EIA) por me fornecer uma estrutura digna para desenvolver meus conhecimentos acerca do mundo da Tecnologia da Informação e aguçar minha curiosidade na constante busca pelo conhecimento.

Em especial às minhas orientadoras, Kate Revoredo e Fernanda Baião, que acreditaram desde o primeiro momento no tema deste trabalho e não mediram esforços para que o mesmo viesse a ser concluído com êxito.

Por fim, e não menos importante, agradeço a Deus por sempre me proteger e tranquilizar, proporcionando refúgio nos momentos de aflição e preparando o caminho adiante para que seja possível trilhar os próximos desafios.

Gustavo Mendonça:

Deixo expressos meus sinceros agradecimentos às seguintes instituições e pessoas, sem as quais o presente trabalho teria sido impossível:

A esta universidade, seu corpo docente, direção e administração que oportunizaram a janela que hoje vislumbro um horizonte superior, eivado pela acendrada confiança no mérito e ética aqui presentes.

Ao Curso de Ciência da Computação da EIA, e às pessoas com quem convivi nesses espaços ao longo desses anos. A experiência de uma produção compartilhada na comunhão com amigos nesses espaços foram a melhor experiência da minha formação acadêmica.

Às minhas orientadoras Kate Revoredo e Fernanda Baião, pelo suporte no pouco tempo que lhe coube, pelas suas correções e incentivos.

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a esperança para seguir. Pai, sua presença significou segurança e certeza de que não estou sozinho nessa caminhada.

Aos meus amigos, pelas alegrias, tristezas e dores compartilhadas. Com vocês, as pausas entre um parágrafo e outro de produção melhora tudo o que tenho produzido na vida.

Agradeço ao mundo por mudar as coisas, por nunca fazê-las serem da mesma forma, pois assim não teríamos o que pesquisar o que descobrir e o que fazer, pois através disto consegui concluir a minha monografia.

E a todos que direta ou indiretamente fizeram parte da minha formação, o meu muito obrigado.

SUMÁRIO

Capítulo 1- Introdução	10
1.1 Motivação.....	10
1.2 Objetivo.....	11
Capítulo 2 - Fundamentação teórica	12
2.1 Mercado Financeiro brasileiro	12
2.1.1 Ação	13
2.2 Descoberta de conhecimento em banco de dados (KDD)	16
2.2.1 Seleção de dados	18
2.2.3 Redução e transformação de dados	18
2.2.4 Mineração de dados	19
2.2.5 Interpretação de resultados e avaliação.....	22
Capítulo 3 – Método para análise de informações financeiras e geração de classificador ...	23
3.1 Escopo do trabalho	23
3.2 Procedimento de manipulação dos dados.....	24
Capítulo 4 – Avaliação da proposta	27
4.1 Selecionar e Transformar Dados.....	27
4.2 Aplicar Análise de Sentimento.....	28
4.2.1 Treinamento.....	28
4.2.2 Busca de Notícias e Adição de nova coluna no Dataset	31
4.3 Minerar Dados.....	34
4.3.1 Pré - Processamento	34
4.3.2 Fase de Classificação	35
4.5 Mining StockTec e aplicação no mundo real	40
4.5.1 Interface e Usabilidade	40
4.5.2 Mining StockTec na estratégia de investimento	43
Capítulo 5 – Conclusões e trabalhos futuros.....	45
5.1 Conhecimento aprendido após aplicação.....	45
5.2 Trabalhos Futuros	46

LISTA DE FIGURAS

Figura 1 - Evolução do índice IBOVESPA desde 1996 até o meio do ano de 2013	12
Figura 2- Tendência de baixa - Topos descendentes.....	14
Figura 3 - Tendência de alta - Fundos ascendentes.	14
Figura 4 - Tendência lateral - Topos e Fundos no mesmo nível.....	15
Figura 5 - Processo de evolução de um dado.	16
Figura 6 - Processo de Descoberta de conhecimento (Han, 2011).	17
Figura 7 - Algoritmo SVM em execução.....	20
Figura 8 - Tela de Abertura do Weka.	21
Figura 11 - Processo de análise e geração de conhecimento.	24
Figura 12 - Notícia relata que investigação continua.....	26
Figura 13 - Fechamento ação PETR4 no dia t com variação positiva.	26
Figura 14 - Primeiras linhas do Dataset inicial.	28
Figura 15 - Estrutura do Dataset de Treinamento	29
Figura 16 - Notícias Disponibilizadas pelo site do Yahoo! Finance	32
Figura 17 - Notícias Disponibilizadas pelo site do ADVFN	32
Figura 18 - Dataset definitivo com a coluna "Sentimento" populada.....	33
Figura 19 - Tela de Pré-processamento do Weka.....	35
Figura 20 - Tela de Classificação do Weka.....	36
Figura 21 - Opções do Classificador SMOreg	37
Figura 22 - Salvamento do modelo gerado pelo Classificador	38
Figura 23 - Carregamento do Modelo	39
Figura 24 - Classificação de uma nova instância	39
Figura 25 - Tela inicial do Mining StockTec.....	40
Figura 26 - Tela de Geração de Base de Dados.....	41
Figura 27 - Tela de Predição.....	42
Figura 28 - Tela de Resultado do Mining StockTec.....	42

LISTA DE ABREVIATURAS

API - Application Programming Interface

BMF&BOVESPA - Bolsa de Valores, Mercadorias e Futuros de São Paulo

CVM – Comissão de Valores Mobiliários

SVM – Support Vector Machine

KDD – Knowledge discovery in databases

RESUMO

Desde que Charles Dow criou o índice Dow Jones em 1896 para compreender o movimento do mercado industrial dos Estados Unidos, busca-se cada vez mais entender o padrão de variação dos preços acionários. Tal feito, ainda é ambição de muitos investidores e instituições financeiras que empregam grandes somas de dinheiro para aprender padrões e aplicá-los no mercado financeiro para angariar lucro em uma transação de compra/venda na bolsa de valores.

Através de técnicas modernas de predição e classificação é possível chegar a resultados, que quando utilizados corretamente geram bons lucros. No mercado financeiro, ganhar dinheiro é a força principal que movimenta todas as transações e analisá-la pode ser o ponto chave para obter mais sucesso na bolsa de valores.

Por outro lado, a mineração de dados é utilizada para a partir de um conjunto de dados aprender um modelo capaz de prever valores futuros. A análise de sentimentos visa entender qual o significado da informação que está sendo utilizada e seu impacto no meio em que está inserida.

Estas duas técnicas estão presentes neste trabalho que propõe a utilização da técnica de mineração de dados aliada à análise de sentimento com o intuito de definir qual é a tendência de uma determinada ação no presente dia de negociação e entender o nível de influência de uma notícia no preço acionário.

O software Stock MiningTec, proposto neste trabalho, é capaz de auxiliar o investidor no momento de tomada de decisão acerca de qual ação vender ou comprar fazendo uso do modelo gerado através da mineração de dados e informações da análise de sentimento.

Palavras-chave: KDD, Análise de sentimento, ações, mercado financeiro, mineração de dados, SVM, investimento

ABSTRACT

Since Charles Dow created the Dow Jones index in 1896 to understand the movement of the industrial market in the United States, investors pursue to acquire the knowledge to discover patterns of stock prices variation. Such accomplishment is still an ambition of many investors and financial institutions that apply large amounts of money in learning and applying methods in the financial market to gather profit in a buy/sell transaction at the stock market.

by applying modern techniques for prediction and classification, it is possible to obtain satisfactory results, that when correctly used generates profit. At the financial market, getting money is the main force that moves every transaction, and its analysis may be the key to obtain success at the stock market.

On the other hand, data mining is used to, from a dataset, learn a model able to predict future values. The sentiment analysis looks to understand the meaning of the information that is being used and the impact in the environment it is inserted in.

These two techniques are present at this study that proposes the use of the data mining combined with the sentiment analysis with the intention to predict what the tendency of a determined stock is at the present negotiation day and understand the influence level of news in the stock price.

The Mining StockTec software proposed at this study is able to help the investor in the decision making about which stock to sell or buy, using the model generated through data mining and sentiment analysis.

Keywords: KDD, Opinion mining, stocks, financial market, Data Mining, SVM, investing

Capítulo 1- INTRODUÇÃO

O objetivo deste capítulo é contextualizar o assunto abordado neste trabalho, bem como apresentar a motivação, seu objetivo e a estrutura dos capítulos.

1.1 Motivação

O mercado de ações é regido pelas leis básicas de oferta e procura e acredita-se que os preços históricos são extremamente importantes. Porém, o movimento do mercado não é definido e se encontra em constante mudança. Como antever as variações do mercado e gerar uma estratégia lucrativa é uma busca incessante.

Com isso, surgiu o emprego de variadas técnicas que visam antecipar a movimentação do mercado para que o investidor tenha como prever as oscilações de mercado que são fruto de diversas variáveis externas. No mercado financeiro, existem várias estratégias diferentes que convergem para um objetivo em comum, o de gerar lucro.

A análise fundamentalista e técnica são abordagens que auxiliam o investidor no momento de análise do posicionamento estratégico em uma empresa específica. A primeira, é uma metodologia baseada na análise dos fundamentos da empresa e da economia, tenta avaliar a saúde financeira da corporação e fazer uma projeção do comportamento futuro de preços das ações. A segunda, estuda o movimento do mercado através de gráficos e indicadores financeiros a fim de prever tendências (Moore, 2002).

Entretanto, nenhuma delas é completamente acurada e possuem diversas vertentes que focam especificamente em um indicador econômico, tornando a estratégia traçada ineficiente.

Graças ao avanço da Tecnologia da Informação, é possível alinhar tais técnicas com as milhares de informações existentes acerca das empresas no meio digital como: rentabilidade, preço acionário, notícias, confiabilidade e outras mais.

Contudo, não existem trabalhos focados no mercado brasileiro envolvendo a influência das notícias e o preço das ações para que investidor seja capaz de tomar a decisão que melhor lhe favorece em relação ao momento corrente do cenário econômico mundial.

1.2 Objetivo

Com a mineração de dados (Vapnik, 1995) é possível a partir de um conjunto de dados aprender um modelo capaz de prever valores futuros e a análise de sentimentos (Liu, 2008), técnica que procura compreender o significado da informação analisada, é capaz de enriquecer este modelo.

Este trabalho de conclusão de curso visa aprender um modelo que seja capaz de prever o valor futuro de uma ação sob a ótica da descoberta de conhecimento em base de dados do ambiente em que se encontram os papéis acionários utilizando-se também da análise de sentimentos das notícias que se encontram disponíveis na rede mundial de computadores.

A abordagem consiste em analisar notícias relativas a uma determinada empresa listada na bolsa de valores brasileira (BM&FBOVESPA) e relacioná-las com os dados disponíveis utilizando análise de sentimento e regressão.

Além do que foi exposto acima, o trabalho procurará responder se estes dados, uma vez combinados com o preço histórico das ações, podem evidenciar padrões de evolução de preços (tendência de alta, estagnação e tendência de baixa) e qual o impacto real dos mesmos no valor final dos papéis acionários que são negociados na bolsa.

O trabalho se encontra estruturado em 5 capítulos. O primeiro, destina-se a expor a motivação e o objetivo deste trabalho de conclusão de curso. O segundo, introduz os temas que são fundamentais para fazer o melhor proveito da informação tratada nas páginas seguintes. O terceiro capítulo, trata da motivação e apresenta a proposta deste trabalho. O quarto capítulo aborda a aplicação do tema proposto utilizando tudo o que foi estudado. E por fim, o quinto capítulo, que é composto pela conclusão e ideias de trabalhos futuros.

Capítulo 2 - FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo são apresentados conceitos importantes e fundamentais para a compreensão do tema proposto. Ao final deste capítulo, o leitor terá conhecimentos dos movimentos financeiros na bolsa de valores, descoberta de conhecimento em base de dados (Fayyad, 1996) e análise de sentimento (Liu, 2008).

2.1 Mercado Financeiro brasileiro

O mercado financeiro, no Brasil, está em pleno processo de crescimento e ampliação como pode ser verificado na figura 1 que demonstra a evolução do principal índice da bolsa de valores, o IBOVESPA, nos últimos anos. A estabilização da economia e a globalização puderam atrair a atenção dos brasileiros ao mercado de ações.

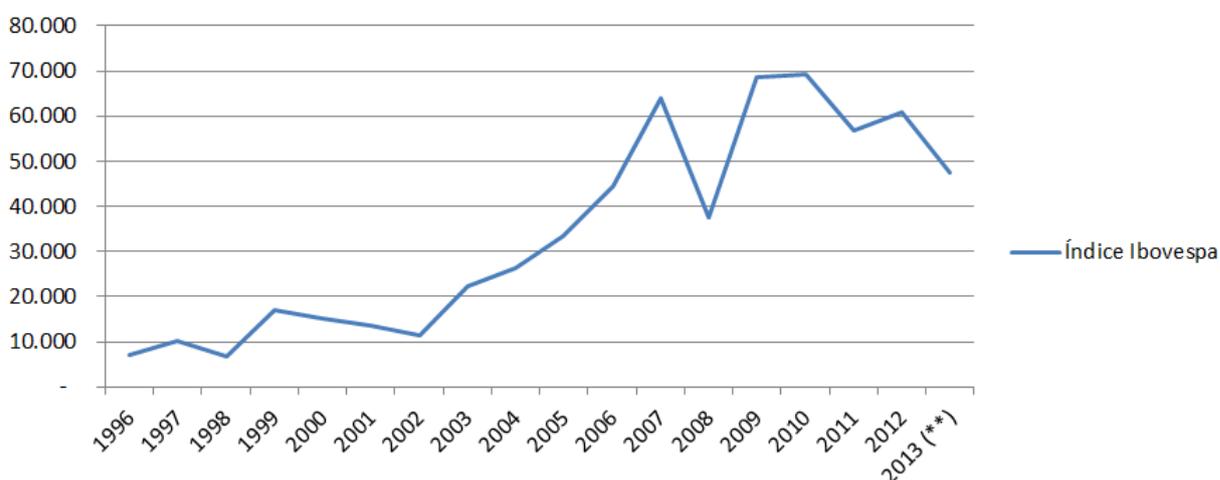


Figura 1 - Evolução do índice IBOVESPA desde 1996 até o meio do ano de 2013 de acordo com o site Minhas Economias.

Com isso, os brasileiros que antes somente utilizavam a poupança como forma de investimento, agora buscam diversos outros mercados para investir a fim de diversificar suas aplicações como a bolsa de valores. O mercado de ações no Brasil é estruturado através da Comissão de Valores Mobiliários (CVM), Companhia

Brasileira de Liquidação e Custódia (CBLIC) e BMF&BOVESPA (Wolwacz, 2002).

A Bolsa é um mercado organizado que mantém um sistema de negociação eletrônico adequado à realização de transações em que são negociadas, entre outras coisas, as ações (Moore, 2002).

A BMF&BOVESPA é uma companhia de capital aberto criada em 2008 a partir da união da Bolsa de Valores de São Paulo e da Bolsa de Mercadorias & Futuros. Atualmente, é o principal mercado de ações da América Latina e um dos maiores do mundo.

Segundo Moore, a Bolsa de Valores ajuda as empresas a se expandirem, gerando assim mais empregos no país, mais receitas de impostos e até mais divisas em moeda estrangeira, no caso de companhias exportadoras. Além disso, com a população investindo em ações ocorre um fortalecimento da saúde financeira, formando assim, um mercado de capitais forte que auxilia no crescimento do país.

2.1.1 Ação

Objeto de estudo e análise deste trabalho, uma ação ou papel, representa parte de uma empresa de capital aberto e consiste na menor parcela do capital social da mesma. O detentor da ação possui uma parcela da empresa e torna-se sócio desta companhia. O acionista pode negociar sua participação em bolsa desde que atento aos critérios regulatórios que envolvem a operação. Seu poder é diretamente proporcional à quantidade e tipo das ações que possui.

As ações se dividem em dois tipos, ordinárias (ON) ou preferenciais (PN) (Moore, 2002). É importante lembrar que essa divisão só ocorre no mercado financeiro do Brasil. As ações ordinárias dão direito a voto nas assembleias dos acionistas, além de participação nos resultados financeiros das companhias. As ações preferenciais dão aos titulares a prioridade na distribuição dos dividendos, ou seja, de parte do lucro da companhia. Em geral, não concedem direito a voto nas assembleias ou dão direito a voto restrito. São, de modo geral, mais negociadas no mercado secundário que as ordinárias. Normalmente, pagam 10% a mais de dividendos, caso o estatuto não defina um dividendo mínimo ou fixo. As ações PN dão também prioridade, no caso de dissolução da empresa, no reembolso de capital.

Uma ação é composta por um nome que identifica a empresa, e possui um valor que varia conforme a percepção do mercado em relação à empresa. A variação do valor de mercado possui 3 (três) tendências quando observados seus topos e fundos. São elas: tendência de lado, tendência de alta e tendência de baixa.



Figura 2- Tendência de baixa - Topos descendentes.

No cenário da primeira tendência a relação entre os topos e fundos é de queda (figura 2). Cada novo topo é mais baixo que o anterior, e cada novo fundo é mais baixo que o anterior. Na tendência de alta, ocorre o inverso: observam-se topos sucessivamente maiores que os anteriores, evidenciando assim, uma alta no valor da ação (figura 3).



Figura 3 - Tendência de alta - Fundos ascendentes.

Em contrapartida, na tendência de lado os topos e fundos variam no mesmo nível, não sendo possível identificar um padrão que seja de queda ou alta (figura 3).



Figura 4 - Tendência lateral - Topos e Fundos no mesmo nível.

Estes três tipos de tendências do mercado de ações são importantes pois são utilizados neste trabalho para relacionar e classificar o tipo da notícia ao papel acionário em análise.

2.2 Descoberta de conhecimento em banco de dados (KDD)

Com o avanço da computação pessoal e poder de processamento, tornou-se possível analisar e compreender o grande volume de informações existente nas base de dados. Para isso, pode ser necessário utilizar um processo denominado descoberta de conhecimento em banco de dados ou KDD, do inglês *knowledge discovery in databases*.

Data mining ou mineração de dados consiste em extrair conhecimento de grande quantidades de dados e apesar de ser utilizado como sinônimo de descoberta de conhecimento, é na verdade, uma parte do processo de descoberta de conhecimento (Fayyad, 1996). Nesta fase, ocorre a aplicação de algoritmos com a finalidade específica de identificar padrões nos dados que sejam válidos, novos, potencialmente úteis e compreensíveis (Fayyad, 1996). Um padrão é definido como um tipo de modelo de uma declaração. Uma instância de um padrão é uma declaração em uma linguagem de alto nível que descreve uma informação preferencialmente interessante, descoberta nos dados de acordo com algum critério estabelecido (Klosgen, 1992).

O KDD é aplicado em uma base de dados que consiste em um conjunto de objetos de dados e seus atributos. É importante compreender a distinção entre dados, informação e conhecimento (figura 5).



Figura 5 - Processo de evolução de um dado.

Um dado é a representação mais bruta, ele possui pouco ou nenhum significado para quem o observa. No que diz respeito à informação, pode-se

entendê-la como um dado mais trabalhado, que pode ser combinada com textos, imagens ou metadados.

Por fim, o conhecimento é a conexão criada entre os diversos pedaços de informação gerando um conjunto organizado e compreensível de tais informações que, através de seu significado, agregarão valor para quem a observa.

O processo de descoberta de conhecimento (figura 6) pode ser dividido nos seguintes passos (Fayyad, 1996) que são abordados a seguir:

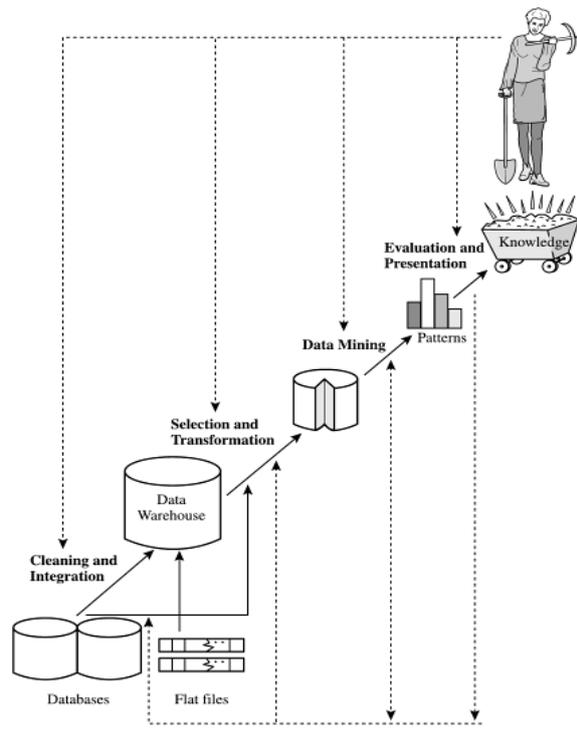


Figura 6 - Processo de Descoberta de conhecimento (Han, 2011).

2.2.1 Seleção de dados

Este é o primeiro passo do processo de KDD: uma carga inicial de dados é selecionada dentre a grande quantidade de dados guardados no banco de dados. A mineração de dados ocorrerá nestes dados. É necessário conhecimento prévio da natureza e relevância dos dados selecionados para a fase de descoberta de conhecimento.

2.2.2 Pré-processamento e limpeza de dados

Os dados do “mundo real” contém ruídos, valores fora do padrão ou em branco, erros e etc. Este passo é necessário para preparar os dados para a fase de mineração. Técnicas de estatísticas são empregadas para tratar a remoção de ruídos e valores perdidos. Conhecimento prévio sobre o domínio da aplicação também pode vir a ser útil nesta fase.

2.2.3 Redução e transformação de dados

Muitos bancos de dados do mundo real são extensos e possuem um grande número de atributos.

Este passo visa selecionar os atributos mais relevantes para a tarefa de mineração de dados, o que consiste na redução dimensional. Por exemplo, pode não ser interessante, dependendo do contexto da análise de informações financeiras, o volume de negociação diário de uma ação.

Em contrapartida, pode ser conveniente visualizar o dado em outra forma, como mudar uma data formatada em 02-02-2014 para 02.02.2014. Tem-se neste caso então, um exemplo de transformação de dados.

2.2.4 Mineração de dados

A fase de mineração de dados consiste em escolher um algoritmo de aprendizado de máquina (Mitchell, 1997) e aplicá-lo em um problema para descobrir padrões. Junto disto, um conhecimento humano acerca do domínio da aplicação também pode ser incorporado para melhorar o processo de mineração de dados.

Os padrões podem ser modelados através de, por exemplo, árvores de decisão ou regras, pontos em um espaço multidimensional (Vapnik,1995), rede Bayesiana (Fayyad, 1996) ou etc.

Atividades típicas da fase de mineração de dados incluem: classificação, clusterização, sumarização, modelo de dependência, detecção de desvio e regressão (Liao, T. W.; Chen, J. H.; Triantaphyllou, E.,1999).

Neste trabalho foram utilizadas classificação e regressão analítica, através da técnica denominada Supporting Vector Machine (SVM), para realizar a fase de mineração de dados (Vapnik, 1995).

O SVM vem sendo aplicado com êxito em várias áreas de predição, como por exemplo, na predição do mercado financeiro (Mukherjee, 1997; e Cao; Tay, 2001), marketing (Ben-david; Lindenbaum, 1997), categorização de textos (Joachims, 2002), detecção de face utilizando imagem (Osuna, 1997) e diagnóstico médico (Tarassenko, 1995).

O SVM é um método de aprendizado supervisionado que analisa dados e reconhece padrões, é usado para classificação e análise de regressão.

O SVM padrão é um classificador binário não-probabilístico, o que significa que ele prevê, para cada entrada, qual de duas possíveis classes esta entrada faz parte. Como o SVM é um classificador, então, dado um conjunto de exemplos de treinamento, cada instância é marcada como pertencente a uma das duas classes, um algoritmo de treino SVM constrói um modelo que prevê se um novo exemplo cai em uma categoria ou outra.

Intuitivamente, um modelo SVM é uma representação de exemplos como pontos no espaço, mapeados para que os exemplos de diferentes categorias sejam divididos por um vão claro que é o mais largo possível. Novos exemplos então são

mapeados dentro deste mesmo espaço e previstos para pertencer a uma categoria baseado em que lado do vão eles caem.

De uma maneira mais formal, um SVM constrói um hiperplano ou um conjunto de hiperplanos em um espaço de alta dimensão ou de dimensão infinita, que pode ser usado para classificação, regressão ou outras tarefas. Intuitivamente, uma boa separação é conseguida pelo hiperplano que tem a maior distância para os pontos de dados de treinamento mais próximos, já que normalmente, quanto maior a margem, menor é o erro de generalização do classificador.

Enquanto o problema original pode ser definido em um espaço dimensional finito, ele muitas vezes ocorre em um espaço no qual os conjuntos a serem discriminados não são separados linearmente. Por isso, foi proposto que o espaço dimensional finito original fosse mapeado em um espaço dimensional muito mais alto, fazendo com que a separação seja mais fácil nesse espaço.

Esquemas SVM usam um mapeamento em um espaço mais largo para que os produtos possam ser computados facilmente em termos das variáveis no espaço original fazendo o tempo de processamento ser razoável (ver figura 7). Para a demonstração matemática completa, consulte (Ou, Phichhang; Wang, Hengshan et al,2009).

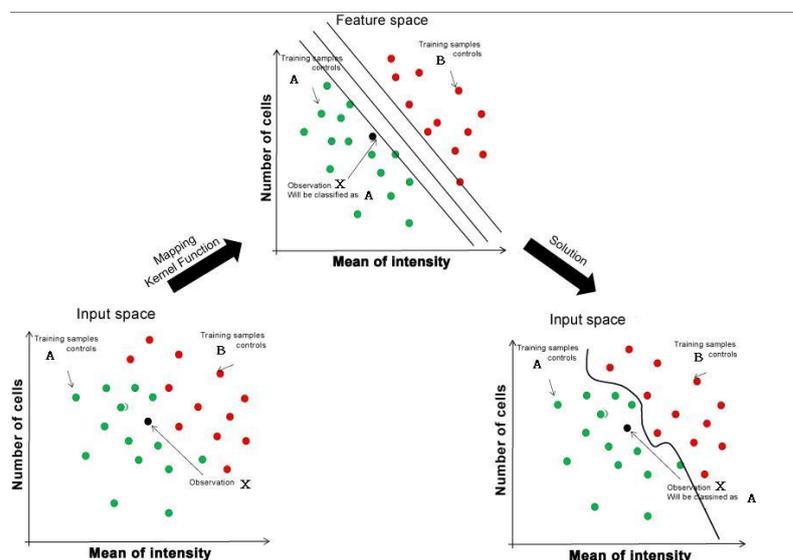


Figura 7 - Algoritmo SVM em execução.

A ferramenta que será utilizada neste estudo é o Weka. O Weka (Waikato Environment for Knowledge Analysis) é uma ferramenta muito popular para

aprendizado de máquina que é escrita em Java e foi desenvolvida na Universidade de Waikato, na Nova Zelândia. O Weka é uma ferramenta livre e é disponibilizado em seu site.

O Weka contém uma coleção de ferramentas de visualização e algoritmos para análise de dados, junto com interfaces gráficas para acesso fácil para essas funcionalidades.

Ele suporta diversas tarefas básicas para a mineração de dados, mais especificamente: clusterização, classificação e regressão.

A tela inicial do Weka pode ser vista na Figura 8.

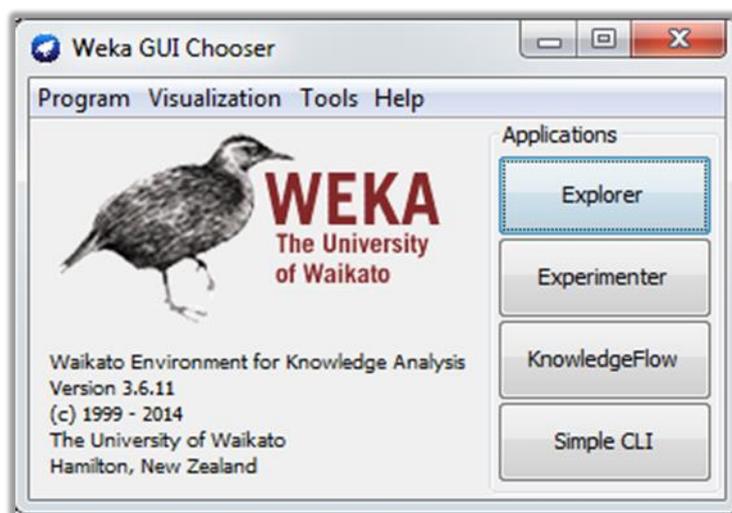


Figura 8 - Tela de Abertura do Weka.

A principal tela do Weka é o Explorer, mas essencialmente as mesmas funcionalidades podem ser acessadas pelas telas de Knowledge Flow e também da linha de comando. Ele também possui o Experimenter, que permite que comparações sistemáticas da performance preditiva da aprendizagem a partir de máquina do Weka em um conjunto de Datasets.

O Weka Explorer possui 6 telas principais, que fazem as tarefas necessárias para a Mineração de Dados:

- Pré-processamento: Esta tela possui métodos para importação de dados de um banco de dados, um arquivo de extensão .csv, etc. e para o pré-processamento destes dados usando algoritmos de filtragem. Esses filtros

podem ser usados para transformar os dados, como por exemplo, transformar atributos numéricos em atributos discretos, e também é possível deletar instâncias e atributos segundo algum critério;

- **Classificação:** Esta tela permite ao usuário aplicar algoritmos de classificação e regressão (chamados indiscriminadamente no Weka de classificadores) ao Dataset resultante, para estimar a acurácia do modelo de predição resultante, e para visualizar predições errôneas, ou o próprio modelo, caso possível;
- **Associação:** Esta tela provê acesso a algoritmos de aprendizado de regras de associação, que tentam identificar todos os interrelacionamentos entre atributos nos dados;
- **Clusterização:** Esta tela dá acesso a técnicas de cluster no Weka, como por exemplo o algoritmo K-Means;
- **Seleção de Atributos:** Esta tela provém algoritmos para identificação dos atributos mais “preditíveis” em um Dataset;
- **Visualização:** Esta tela apresenta uma matriz de gráficos de dispersão, em que gráficos de dispersão individuais podem ser selecionados, além de analisados usando vários operadores de seleção;

2.2.5 Interpretação de resultados e avaliação

Neste passo, os padrões encontrados são interpretados e seus méritos são verificados. De acordo com Fayyad (1996), o primeiro objetivo da mineração de dados é a predição e descrição.

Na predição, pode-se prever ou estimar os valores de variáveis chave a partir de valores de outras variáveis, ou de valores históricos conhecidos sobre a variável chave. Este trabalho é um ótimo exemplo, pois utiliza a predição para tratar a informação acerca dos valores de ações.

Por outro lado, pode-se desejar encontrar descrições e abstrações de grandes quantidades de dados (fluxo total de negociação de dólares durante um dia da bolsa de valores) para um melhor entendimento do cenário.

Capítulo 3 – MÉTODO PARA ANÁLISE DE INFORMAÇÕES FINANCEIRAS E GERAÇÃO DE CLASSIFICADOR

Este capítulo visa introduzir uma forma de selecionar e tratar os dados econômicos utilizando a análise de sentimento com intuito de prever o valor de um papel acionário. Para este fim, o trabalho cria o sistema Mining StockTec que auxiliará o investidor no momento da tomada de decisão e construção da estratégia de investimento.

3.1 Escopo do trabalho

Levando em consideração o que foi discutido nos capítulos anteriores, este trabalho aborda uma análise diferenciada de tratamento dos dados financeiros. O processo descrito na figura a seguir, define uma forma para analisar as informações e utilizar-se dos recursos de processamento de KDD para chegar a um padrão útil que introduza um novo conhecimento acerca dos dados acionários selecionados inicialmente.

Em adição a proposta de geração de conhecimento descrita na figura 11, este trabalho introduz um sistema de apoio à análise financeira denominado Mining Stocktec que utiliza os conhecimentos aprendidos através do KDD para criar uma plataforma útil aos investidores, trazendo informação sobre a tendência da ação em um determinado dia de negociação.

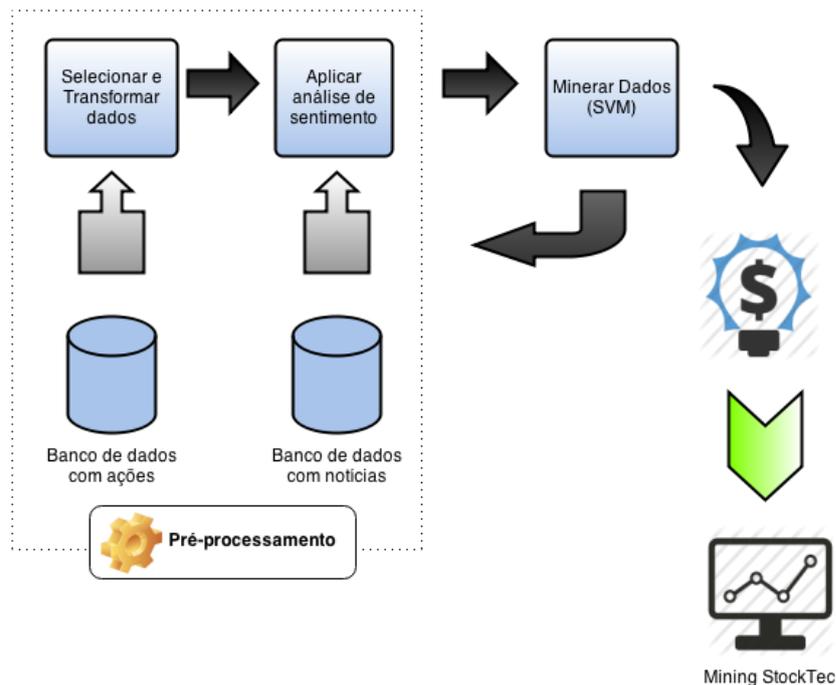


Figura 9 - Processo de análise e geração de conhecimento.

3.2 Procedimento de manipulação dos dados

Para iniciar o processo é preciso dispor de uma base de dados confiável. Além de conter dados consistentes, ela deve conter preferencialmente o maior número possível de informações úteis que represente fielmente o cenário de análise.

A base de dados pode estar disponível na forma de um banco de dados relacional, planilha local, ou armazenado na nuvem. Independente de qual seja a fonte, os atributos essenciais que a base de dados deve conter são: data, os preços acionários de fechamento, abertura, máximo e mínimo.

Para uma maior precisão, é recomendável também que se incluam na base de dados a cotação de Compra e Venda do Dólar em relação ao Real e o Preço de Fechamento do índice S&P 500, que é baseado nas capitalizações de mercados de 500 grandes companhias que têm suas ações listadas na NYSE ou na NASDAQ. Além disto, é necessário definir um período de tempo finito no momento de seleção dos dados, o ideal é que este período compreenda no mínimo 5 (cinco) anos de negociação, porque é um período razoável para que vários cenários diferentes tenham ocorrido no mercado. Sem estes atributos essenciais e um período de

negociação bem definido, não será possível treinar o algoritmo para encontrar um padrão satisfatório.

É essencial a utilização de uma fonte de notícias consistente e que possua imparcialidade quando se tratando de informar acerca de um evento positivo ou negativo que afeta diretamente uma ação listada em bolsa. Para isso, é aconselhável utilizar sites de notícias respeitados no meio financeiro ou até mesmo pagar por acesso a um banco de dados particular, visto que não é trivial encontrar de forma gratuita notícias históricas de ações organizadas temporalmente armazenadas em um só lugar.

No passo seguinte, ocorre uma avaliação destes dados visando assegurar que todo o conteúdo selecionado não possui valores estranhos à semântica do contexto. Por exemplo, nesta fase é possível identificar sem muito esforço o valor string 'Not available' para o campo de preço de abertura que é estritamente numérico.

Uma vez tendo ocorrido o tratamento inicial dos dados o próximo passo é a aplicação da análise de sentimento como visto na seção 2.3. Neste passo, busca-se entender se o significado das notícias selecionadas previamente é positivo ou negativo. Para melhor acurácia, é necessário dispor de um mecanismo voltado para análise de sentimento que foi treinado com palavras similares as do meio financeiro.

É também nesta fase, que deve ser feito o relacionamento entre o dia da notícia, seu significado e ação. Saber se em um determinado dia a ação possui uma notícia boa, ruim ou neutra é essencial para a análise e deve ser adicionado aos dados acionários utilizando valores inteiros como, por exemplo, positivo (1), negativo (-1) e indiferente (0). O cálculo funciona da seguinte maneira, para cada dia, caso este dia tenha mais notícias positivas relacionadas à ação do que negativas, a coluna referente a este dia receberá o valor "1". Caso contrário, receberá o valor "-1". Já caso a quantidade de notícias positivas seja a mesma de negativas, a coluna receberá o valor "0".

O próximo passo é o da mineração de dados com a utilização do algoritmo de SVM. Nesta fase, é utilizada a ferramenta Weka. Também podem ser utilizadas outras ferramentas, como o R, IBM SPSS e RapidMiner.

Com o conhecimento criado ao final do processo anterior é possível, por exemplo, alimentar o programa Mining StockTec para saber se a ação da empresa Petrobras (PETR4) no dia t irá apresentar deslocamento de alta, lateral ou baixa. Para aplicar o conhecimento aprendido, basta selecionar uma notícia do dia anterior $t-1$ conforme a figura a seguir:

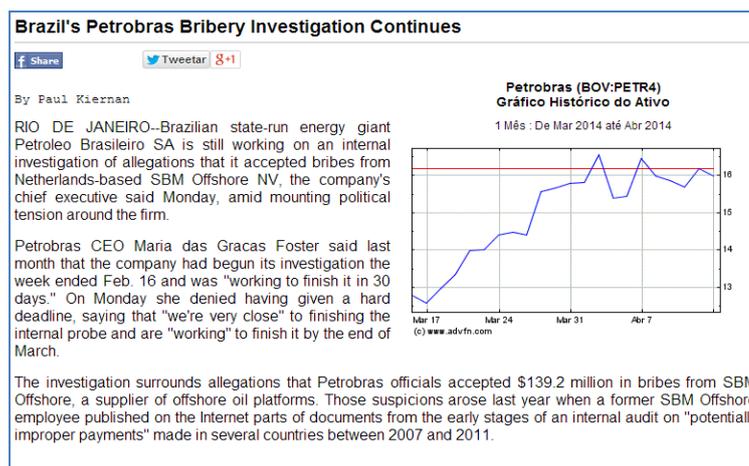


Figura 10 - Notícia relata que investigação continua.

Esta notícia informa ao leitor que a investigação dos casos de suspeita de pagamento irregulares de empresas participantes de licitações para empregados da Petrobras continuam. Além disso, relata que uma comissão investigativa será instaurada no Congresso para analisar o caso e punir os culpados.

Tendo o sistema compreendido esta notícia como positiva através da análise de sentimento, este resultado será utilizado no modelo resultante do processo de KDD para ajustar o resultado da predição. Ao colocar os atributos financeiros e informação sobre a notícia do dia $t-1$, o sistema deve indicar uma provável alta do papel que se assemelhe aos dados reais (ver figura 11) visto a recuperação de valor do papel observado na semana e um cenário de notícia positiva.

Data	Histórico	Var.Dia(%)	Abertura	Mínimo	Médio	Máximo	Volume	Negócios	
18/03/2014	12,97	12,19	+3,18	11,88	11,81	12,06	12,26	542,14M	48.043

Legenda: K - Mil | M - Milhão | B - Bilhão

Figura 11 - Fechamento ação PETR4 no dia t com variação positiva.

Capítulo 4 – AVALIAÇÃO DA PROPOSTA

A proposta deste capítulo é a aplicação do que foi apresentado anteriormente em um estudo de caso real. Para este fim, será utilizada a empresa VALE S.A. Espera-se ao fim do processo, ser possível prever o preço de fechamento da ação da empresa em um determinado dia.

4.1 Selecionar e Transformar Dados

Para aplicação do estudo em um caso real, utilizaremos a empresa VALE S.A., com o propósito de prever o preço de fechamento da ação da empresa em um determinado dia e determinar qual o nível de influência que as notícias relacionadas a corporação implicam.

Para isso, o processo deve começar com a construção da base de dados que será enviada para que o programa possa ler os dados relevantes da empresa em um determinado período de tempo e possa então gerar um modelo que prevê resultados futuros a partir destes dados.

Em seguida, uma nova coluna é inserida na base de dados através da Análise de Sentimentos de notícias relacionadas à ação no mesmo período de tempo em que os dados foram obtidos.

Então, através do Weka, é feita a mineração em cima desses dados, em que pelo algoritmo SVM se gera um modelo que tem como objetivo prever o preço de encerramento desta ação em um dia através dos dados da mesma no dia anterior.

A construção da base de dados inicial começa através da busca pelos dados históricos da ação escolhida. No site do Yahoo! Finance é possível obter os dados históricos da maioria das ações do mercado financeiro em uma planilha¹, com os preços das ações corrigidos com dividendos, e foi através dele que os dados da VALE foram obtidos, sendo eles: Data, Preço de Abertura, Preço de Fechamento, Preço Máximo, Preço Mínimo, Volume (em dólares), Preço de Fechamento Ajustado. Para o estudo, foram excluídas algumas colunas que não serão necessárias.

¹ <http://finance.yahoo.com/q/hp?s=VALE>

Além disso, alguns ajustes tiveram que ser feitos para que os dados se adequassem à proposta, já que eram necessários os dados do dia anterior na mesma linha do Preço de Fechamento da ação no dia.

O histórico das cotações do dólar foi obtido através do site do Banco Central do Brasil, que disponibiliza essas informações em uma planilha². O Preço de Fechamento do S&P 500 também foi obtido através do site do Yahoo! Finance, do mesmo modo que foram extraídos os dados da VALE³.

Finalmente, a base de dados Inicial contém todos os atributos essenciais descritos no Capítulo 3, além daqueles que eram considerados recomendáveis. As primeiras linhas da base de dados podem ser vistas na Figura 12.

Date	Open Anterior	High Anterior	Low Anterior	Close	Close Anterior	Volume Anterior	Dolar Compra Anterior	Dolar Venda Anterior	S&P500 Close Anterior
25/04/2014	13.71	14.05	13.59	13.51	13.86	19982700	2.2223	2.2229	1878.61
24/04/2014	13.62	13.65	13.44	13.86	13.59	11744700	2.242	2.2426	1875.39
23/04/2014	13.78	13.86	13.67	13.59	13.68	17194900	2.2443	2.2449	1879.55
22/04/2014	14.24	14.24	13.68	13.68	13.8	16784300	2.2476	2.2482	1871.89
21/04/2014	13.89	14.24	13.81	13.8	14.1	14015800	2.2476	2.2482	1864.85
17/04/2014	14.04	14.06	13.79	14.1	13.87	14325800	2.2336	2.2342	1862.31

Figura 12 - Primeiras linhas do Dataset inicial.

4.2 Aplicar Análise de Sentimento

4.2.1 Treinamento

Para gerar um modelo que possa classificar uma notícia como sendo positiva ou negativa, deve-se fazer um treinamento a partir de um Dataset contendo tanto notícias boas quanto notícias ruins, cada uma com sua devida classificação.

Desta maneira, após o treinamento feito, o modelo será gerado e então será possível classificar uma notícia nova como sendo positiva ou negativa. Este é o objetivo principal da Análise de Sentimentos neste trabalho.

² <http://www4.bcb.gov.br/pec/taxas/port/ptaxnpesq.asp?id=txcotacao>

³ <http://finance.yahoo.com/q/hp?s=%5EGSPC+Historical+Prices>

O Dataset utilizado neste estudo de caso, foi o construído por Richard Johansson⁴. Este Dataset possui reviews de produtos da Amazon, organizados por Categorias (Livros, Câmeras, DVD, Saúde, Música, Software) e Classificação (Positivo, Negativo) e o processamento foi realizado dentro do Mining StockTec.

A estrutura dele é feita com as pastas das Categorias, sendo que dentro de cada uma delas existem as páginas de Positivo e Negativo, para separar as notícias. Já as notícias, são arquivos de extensão .txt contendo o texto da notícia correspondente à categoria na qual ela pertence.

A estrutura pode ser vista pela figura 13. Para este estudo de caso, a classificação por Categoria não fazia sentido, portanto, ele foi reorganizado de modo em que os reviews eram apenas organizados por Classificação.

books	4/27/2014 6:46 PM	Pasta de arquivos
camera	4/27/2014 6:46 PM	Pasta de arquivos
dvd	4/27/2014 6:47 PM	Pasta de arquivos
health	4/27/2014 6:47 PM	Pasta de arquivos
music	4/27/2014 6:47 PM	Pasta de arquivos
software	4/27/2014 6:47 PM	Pasta de arquivos

neg	4/27/2014 6:46 PM	Pasta de arquivos
pos	4/27/2014 6:46 PM	Pasta de arquivos

1	2/14/2012 7:39 AM	Documento de Te...
1_2	2/14/2012 7:39 AM	Documento de Te...
1_3	2/14/2012 7:39 AM	Documento de Te...
1_4	2/14/2012 7:39 AM	Documento de Te...
1_5	2/14/2012 7:39 AM	Documento de Te...

Figura 13 - Estrutura do Dataset de Treinamento

Como este banco de dados é de um domínio diferente do proposto neste estudo, teve que ser feito um teste para que se houvesse a certeza de que o classificador gerado a partir do treinamento deste banco seria capaz de classificar textos de domínios diferentes deste.

⁴ <http://www.cse.chalmers.se/edu/year/2013/course/TIN171/amazon-balanced-6cats.zip>

Para isso, foram fornecidas 50 notícias para que ele classificasse como positivas ou negativas, sendo que estas mesmas notícias seriam classificadas manualmente para que a comparação pudesse ser feita.

Das 50 notícias, 41 tiveram o mesmo resultado na comparação do manual com o do classificador. Logo, a conclusão foi que este classificador estava apto para classificar as notícias deste estudo, pois obteve um nível aceitável de acerto para o que foi proposto no trabalho.

O algoritmo escolhido para fazer o treinamento do Dataset foi o algoritmo de Análise de Sentimento do Lingpipe⁵.

O Lingpipe é uma biblioteca JAVA, disponibilizada online, que possui ferramentas de processamento de texto, que são usadas através de linguísticas computacionais. Ele é usado para tarefas como a procura de nomes de pessoas, organizações ou lugares dentro de um texto, classificar resultados de pesquisa do Twitter automaticamente em categorias e sugerir correções em consultas. Neste caso, ele será usado para fazer a análise de sentimentos das notícias da ação e classificá-las como positivas ou negativas.

Para que o algoritmo seja treinado, é necessário que se crie uma classe que possa ler o Dataset elaborado anteriormente. Neste caso, esta classe se chama “PolarityBasic”. Então, um novo classificador é gerado a partir da criação de uma instância desta classe.

Este classificador, a partir de um endereço passado no construtor da classe, trata cada pasta contida neste endereço como uma classificação, e todos os arquivos dentro desta pasta como instância.

No caso deste estudo de caso, foi passado o endereço contendo as pastas “pos” (positiva) e “neg” (negativa) e dentro destas pastas, os arquivos contendo as notícias relacionadas a cada classificação.

Assim, o algoritmo é treinado e gera um modelo que é guardado dentro do classificador.

A partir disto, com este classificador, já é possível classificar qualquer notícia que seja passada a ele que tenha associação e que seja da mesma língua que as notícias do Dataset usado para treinamento do algoritmo, como Positiva ou Negativa. Por exemplo, o texto abaixo é classificado como uma notícia positiva.

⁵ <http://alias-i.com/lingpipe/>

“Brazilian mining major Vale SA said Tuesday it has secured a 6.2 billion real (\$2.8 billion) financing package from state development bank BNDES for its massive iron-ore project in the Carajas region of the Amazon jungle. Vale should receive the funds within three years and will have 10 years to pay them back. Additional terms of financing weren't disclosed. The new contract comes in addition to some BRL12 billion in credit lines and financing packages that Vale already has with BNDES, the company said. The world's largest producer of iron ore plans to invest \$8.09 billion opening a new mining block at Carajas and another \$11.58 billion in logistics to expand its production capacity in the region by 90 million metric tons through 2018. Vale last year churned out 311 million tons of iron ore across Brazil.”

Como é possível visualizar, nesta notícia o veículo de informação afirma que a empresa VALE contraiu um grande empréstimo para aplicação em um grande projeto e trata da expansão prevista para os próximos anos.

4.2.2 Busca de Notícias e Adição de nova coluna no Dataset

Para que a análise de sentimento das notícias relacionadas à ação entrasse no Dataset, foi criada uma nova coluna, chamada de “Sentimento”. No primeiro momento, todas as linhas tiveram o valor dessa coluna setado como 0.

O próximo passo era achar uma boa base de dados que tivesse notícias sobre ações no mercado financeiro para que as mesmas pudessem ser extraídas e classificadas, e assim, a nova coluna do Dataset fosse populada.

As opções eram o site do Yahoo! Finance⁶ e o site do ADVFN⁷, que são grandes portais relacionados a mercado financeiro e ações.

O Yahoo! Finance é um site patrocinado pelo Yahoo! que provê informações financeiras e comentários sobre o mercado financeiro. O site oferece informações como preços de ações e notícias sobre uma ação, como pode ser visto na figura 14, que trás as últimas notícias sobre a VALE.

⁶ <https://finance.yahoo.com/>

⁷ www.advfn.com

Thursday, April 24, 2014	
•	Brazil court rules in Vale's favor in foreign-taxation dispute Reuters (Thu, Apr 24)
•	BHP Billiton Ltd Is an Efficiency Machine at Motley Fool (Thu, Apr 24)
Wednesday, April 23, 2014	
•	Cliffs Natural Resources, Inc. Earnings: Is the Bottom Finally In? at Motley Fool (Wed, Apr 23)
•	Vale SA Likely to Lose Its Guinea Investment at Motley Fool (Wed, Apr 23)
Tuesday, April 22, 2014	
•	3 Basic Materials Stocks Dragging The Sector Down at TheStreet (Tue, Apr 22)

Figura 14 - Notícias Disponibilizadas pelo site do Yahoo! Finance

O ADVFN é um site sobre mercado financeiro, que provêm dados e serviços para investidores privados como preços de ações, tabelas, notícias, e outras informações para cada empresa que tenha ações nas principais bolsas do mundo. Na figura 15 podem ser vistas as primeiras notícias da seção da VALE no site.

Data	Hora	Fonte	Título
25/04/2014	09:40	DJN	Brazil Court Rules in Favor of Vale
25/04/2014	09:00	DJN	Brazilian Court Rules in Favor of Vale in Foreign-Tax Dispute
15/04/2014	18:02	DJN	ADR Shares End Lower: Infosys, Diageo Shares Active
15/04/2014	14:06	DJN	Brazil's Vale Secures Loan Package From BNDES -- Update
15/04/2014	11:02	DJN	Brazil's Vale Secures BRL6.2 Billion Loan Package From BNDES

Figura 15 - Notícias Disponibilizadas pelo site do ADVFN

Neste trabalho, o banco de dados escolhido para que as notícias fossem buscados foi o do ADVFN. Ele foi escolhido porque o site do Yahoo! Finance traz apenas notícias de um pequeno período de tempo até hoje, não disponibilizando as notícias mais antigas de seu banco de dados pelo site. Ele também dá a possibilidade de baixar um arquivo XML com o Feed RSS⁸ das notícias, porém o Feed também só retorna as últimas notícias da ação, o que não é de interesse neste trabalho.

⁸ A tecnologia do RSS permite aos usuários da internet se inscreverem em sites que fornecem "feeds" RSS. Estes são tipicamente sites que mudam ou atualizam o seu conteúdo regularmente. Para isso, são utilizados Feeds RSS que recebem estas atualizações, desta maneira o utilizador pode permanecer informado de diversas atualizações em diversos sites sem precisar visitá-los um a um.

O ADVFN, por outro lado, traz as notícias de um longo período de tempo até hoje para usuários registrados, que é aquilo que é preciso para a sequência no estudo de caso.

A partir da tabela da Figura 15, acessando cada link de notícia, é possível obter o texto de notícias históricas relacionadas a essa ação.

Para que essas notícias fossem buscadas de modo automático, de forma que então o software pudesse classificá-las, foi necessário o uso de uma biblioteca JAVA chamada de JSOUP⁹. O JSOUP tem o intuito de fazer o download de páginas web e extrair elementos HTTP desta página.

Com a ajuda do JSOUP, foi criada uma função que loga no site através de um POST, e então extrai cada linha da tabela da seção de notícias da VALE. A partir daí, o texto desta notícia é classificado através do algoritmo que foi treinado anteriormente.

A Análise de Sentimento é incluída na tabela de acordo como foi explicado anteriormente no Capítulo 3.

Assim, depois de todas as notícias terem sido classificadas, a coluna “Sentimento” foi devidamente populada e o Dataset passa a ser definitivo. As primeiras linhas do novo Dataset podem ser vistas na Figura 16.

Date	Open Anterior	High Anterior	Low Anterior	Close	Close Anterior	Volume Anterior	Dolar Compra Anterior	Dolar Venda Anterior	S&P500 Close Anterior	Sentimento
25/04/2014	13.71	14.05	13.59	13.51	13.86	19982700	2.2223	2.2229	1878.61	-1
24/04/2014	13.62	13.65	13.44	13.86	13.59	11744700	2.242	2.2426	1875.39	0
23/04/2014	13.78	13.86	13.67	13.59	13.68	17194900	2.2443	2.2449	1879.55	0
22/04/2014	14.24	14.24	13.68	13.68	13.8	16784300	2.2476	2.2482	1871.89	0
21/04/2014	13.89	14.24	13.81	13.8	14.1	14015800	2.2476	2.2482	1864.85	0
17/04/2014	14.04	14.06	13.79	14.1	13.87	14325800	2.2336	2.2342	1862.31	0
16/04/2014	14.24	14.25	13.5	13.87	13.9	32329600	2.2251	2.2257	1842.98	0
15/04/2014	14.9	15	14.76	13.9	14.84	15550900	2.209	2.2096	1830.61	1
14/04/2014	14.62	14.82	14.54	14.84	14.76	24240700	2.2053	2.2059	1815.69	0
11/04/2014	14.93	15	14.75	14.76	14.81	20784400	2.1982	2.1987	1833.08	1
10/04/2014	14.88	15.17	14.75	14.81	14.99	23526300	2.2105	2.2111	1872.18	0
09/04/2014	15.28	15.59	14.98	14.99	15.07	40417900	2.1968	2.1974	1851.96	-1
08/04/2014	14.52	15.05	14.52	15.07	14.98	32978400	2.232	2.2326	1845.04	1

Figura 16 - Dataset definitivo com a coluna “Sentimento” populada.

⁹ www.jsoup.org/

4.3 Minerar Dados

Seguindo com a aplicação do processo apresentado no capítulo 3 e com a base de dados pronta, já é possível avançar para a próxima etapa deste estudo de caso, que é a Mineração de Dados. Esta é a principal etapa do trabalho, já que dela sairá o modelo para predição de preços dessa ação.

A seguir, serão detalhadas as telas que serão usadas neste estudo de caso, a tela de pré-processamento e a tela de Classificação.

4.3.1 Pré - Processamento

A tela de Pré-processamento do Weka pode ser vista na Figura 17. Nesta área, o programa recebe o Dataset no qual ele vai trabalhar em cima. Através do botão “Open file..”, o usuário pode escolher um arquivo de extensão .arff ou de extensão .csv válido para que possa ser usado como Dataset. No caso deste estudo, foi passado o .csv correspondente à base de dados que foi criada nas seções anteriores.

Então, o Weka carrega a base de dados e disponibiliza seus dados para o pré-processamento.

Neste estudo de caso, quase todo o pré-processamento já foi feito manualmente fora do Weka. Assim, o único pré-processamento a ser feito no Weka é a exclusão do atributo de Data, que não vai ser utilizado para a classificação, pois o algoritmo SVM que será utilizado para a classificação não consegue identificar datas.

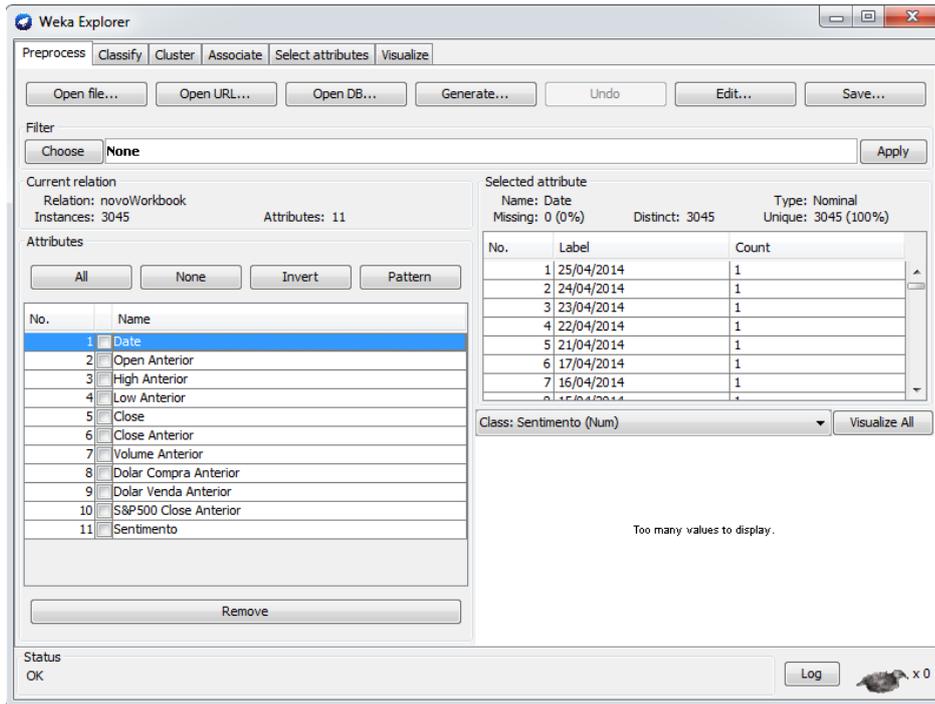


Figura 17 - Tela de Pré-processamento do Weka

4.3.2 Fase de Classificação

A etapa de Classificação é a responsável por gerar o modelo de predição que será utilizado no estudo. A tela de Classificação do Weka pode ser vista na Figura 18.

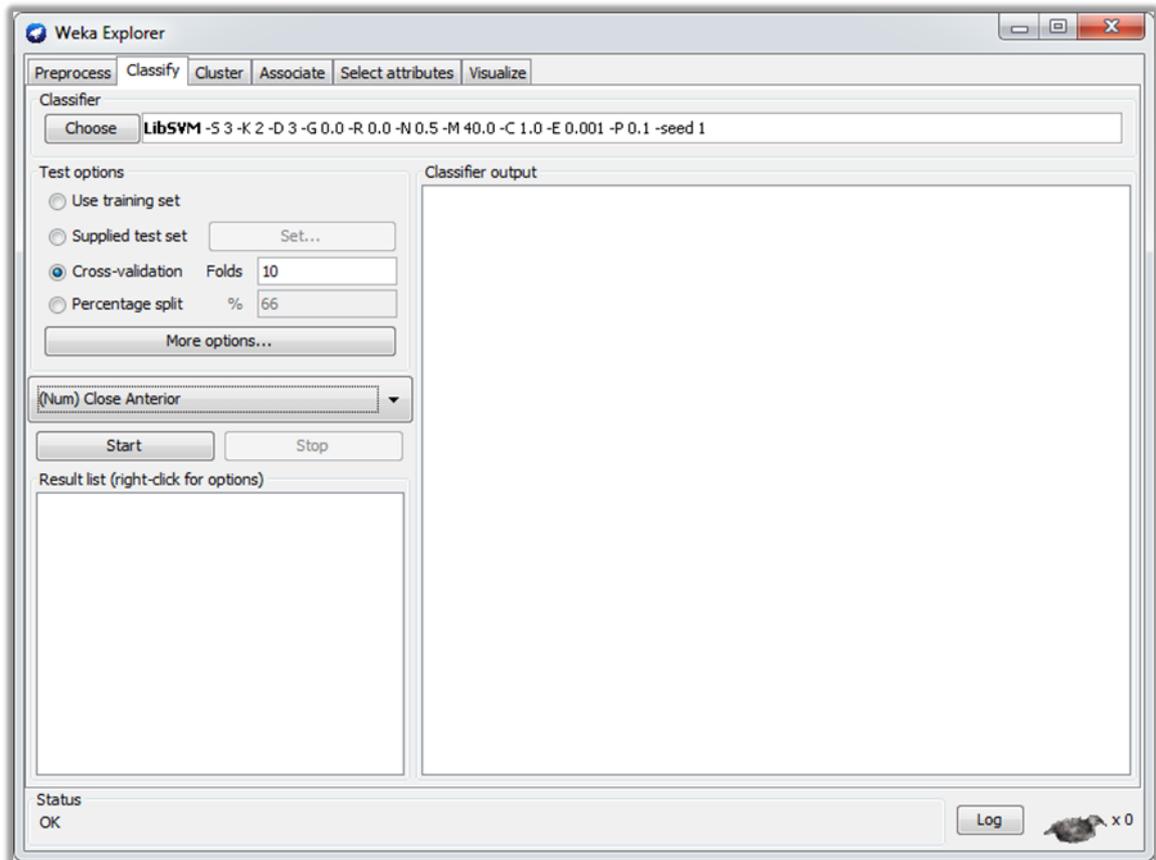


Figura 18 - Tela de Classificação do Weka

O primeiro passo é selecionar o classificador a ser utilizado através do botão “Choose”. Neste caso, o classificador a ser utilizado é o SMOreg. Clicando na caixa de texto ao lado do botão, é possível escolher as opções deste classificador. As opções escolhidas para o estudo estão presentes na Figura 19.

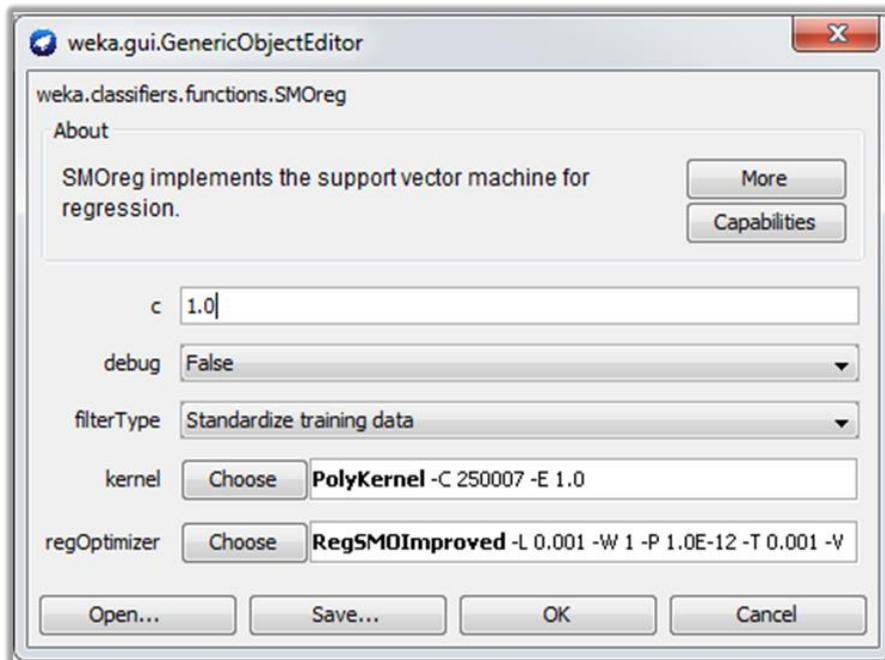


Figura 19 - Opções do Classificador SMOreg

O próximo passo é definir qual será o método de teste utilizado para o modelo que será gerado. Existem 4 possíveis opções:

- “Use Training Set”: O classificador é avaliado em quão bem ele prevê a classe das instâncias em que ele foi treinado;
- “Supplied Test Set”: O classificador é avaliado em quão bem ele prevê a classe de um conjunto de instâncias carregadas de um arquivo, que é escolhido pelo botão “Set...”;
- “Cross-validation” : O classificador é avaliado por cross-validation, usando o número de folds (iterações) que são inseridos no campo de texto “Folds”;
- “Percentage Split” : O classificador é avaliado em quão bem ele prevê uma certa porcentagem dos dados nos quais foram separados para teste. A quantidade de dados separados depende do valor que é inserido no campo “%”.

O método escolhido pelo estudo foi o de Cross-validation, porque ele permite uma maior quantidade de dados a serem testados, fazendo assim com que o modelo tenha uma precisão de erro menor.

O Weka utiliza um método de cross-validation chamado k-fold. Este método consiste na divisão do conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste

e os k-1 restantes são usados para estimação dos parâmetros e calcula-se a acurácia do modelo. Este processo é realizado k vezes alternando de forma circular o subconjunto de teste.

Em seguida, deve ser escolhida a classe, que é o atributo que será predito pelo modelo gerado. Ele é escolhido pelo drop-down localizado acima do botão “Start”.

Depois de toda essa configuração, já é possível gerar o modelo desejado para predição do Preço de Fechamento da ação, dados os seus dados do dia anterior.

Para gerar o modelo, basta clicar no botão “Start”, e então o algoritmo começará a ser rodado.

Após o processamento, o modelo estará disponível a partir da “Result List”. Assim, ele pode ser salvo para ser utilizado para a predição dos valores. O salvamento é feito através da opção “Save Model”, que é habilitada após o usuário clicar com o botão direito do mouse no respectivo modelo na “Results List”, como pode ser visto na Figura 20.

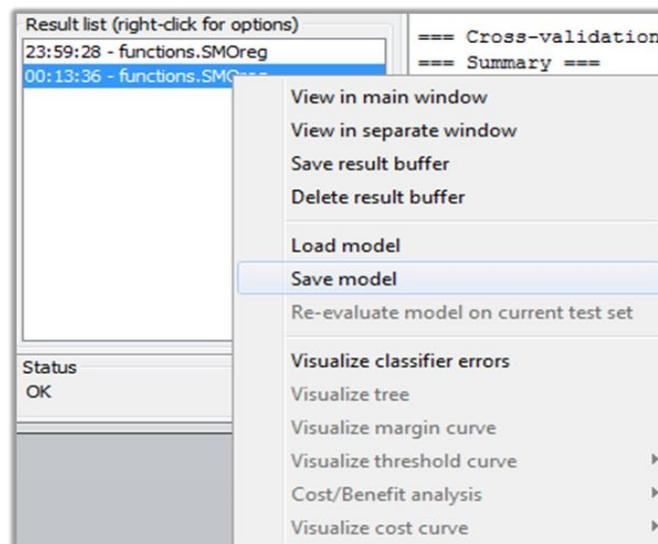


Figura 20 - Salvamento do modelo gerado pelo Classificador

Ao final deste processo, será gerada uma função de predição com base nas instâncias processadas que será utilizada posteriormente no Mining StockTec. Seguindo o estudo de casos, a função retornada neste estágio é:

$$\text{Fechamento} = -0.2149 * (\text{normalized}) \text{ Open Anterior} + 0.1914 * (\text{normalized}) \text{ High Anterior} + 0.1769 * (\text{normalized}) \text{ Low Anterior} + 0.8423 * (\text{normalized}) \text{ Close Anterior} + 0.0005 * (\text{normalized}) \text{ Volume Anterior} - 0.0007 * (\text{normalized}) \text{ Dolar Compra Anterior} + 0.001 * (\text{normalized}) \text{ Dolar Venda Anterior} - 0.0015 * (\text{normalized}) \text{ S\&P500 Close Anterior} - 0.0016 * (\text{normalized}) \text{ Sentimento} + 0.0036$$

4.4 Carregamento do Modelo e Predição

Com o modelo salvo, já é possível aplicar a predição desejada neste estudo de caso. Para isso, é necessário usar a API do Weka para Java. Esta API traz a grande maioria das funcionalidades do Weka para uma biblioteca JAVA que pode ser chamada por um programa para que ele tenha acesso a essas funcionalidades.

Todo o pré-processamento e a classificação poderiam ter sido feitos através do uso de funções da API, mas depois de testes de viabilidade, chegou-se a conclusão de que a classificação ocorreu de modo muito mais demorado através da API. Logo, foi decidido utilizar o próprio Weka separadamente para a classificação.

Porém, para a predição, não é necessário que se se faça pelo Weka, porque este processo é bem rápido, depois que o modelo já está pronto.

O próximo passo é carregar o modelo gerado e salvo pelo Weka para o programa. Isso é feito através da função `read`, que lê o modelo de um arquivo, que é passado como parâmetro na função e retorna um classificador com aquele modelo, como pode ser visto na Figura 21.

```
Classifier cls = (Classifier) weka.core.SerializationHelper.read("/some/where/svm.model");
```

Figura 21 - Carregamento do Modelo

Para fazer a predição de uma instância, é só criá-la e chamar a função `classifyInstance` do classificador gerado. Este método retorna uma String que com a predição desejada. Esta chamada pode ser vista na Figura 22.

```
String resultado = cls.classifyInstance(newInstance);
```

Figura 22 - Classificação de uma nova instância

4.5 Mining StockTec e aplicação no mundo real

4.5.1 Interface e Usabilidade

O sistema foi idealizado para possuir uma interface minimalista que fosse capaz de ser facilmente utilizada. Logo, a tela inicial do Mining StockTec possui dois botões: Gerar Base de Dados e Predizer Ação. A tela pode ser vista na Figura 23.



Figura 23 - Tela inicial do Mining StockTec

Para fazer a adição da coluna de “Sentimento” na base de dados, deve ser escolhida a opção “Gerar Base de Dados”. Nesta tela, vista na Figura 24, o usuário passa uma base de dados com os atributos essenciais descritos no Cap.3, em formato *.xls, para que o sistema possa adicionar a nova coluna à ela através de Análise de Sentimento.

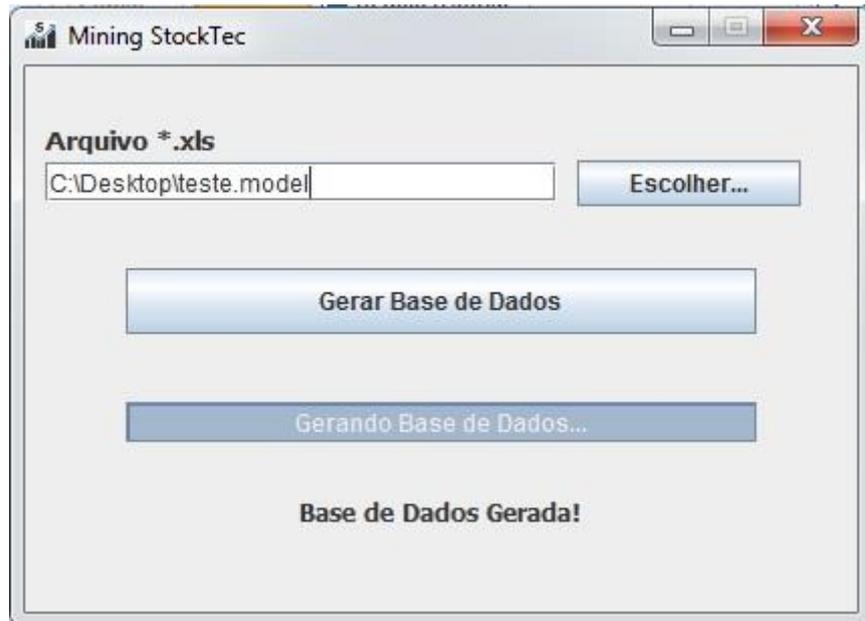


Figura 24 - Tela de Geração de Base de Dados

Caso o modelo já esteja pronto e o usuário quiser prever o preço da ação em algum dia, deve-se escolher a opção "Prever Ação". Nesta tela, vista na Figura 25, deve-se carregar o modelo, que deve estar em formato *.model, e também inserir as informações necessárias para que o modelo possa retornar o preço que foi calculado.

A tela do resultado pode ser vista na Figura 26. Caso a ação tenha uma previsão de alta, o texto aparecerá em verde. Caso contrário, aparecerá em vermelho.

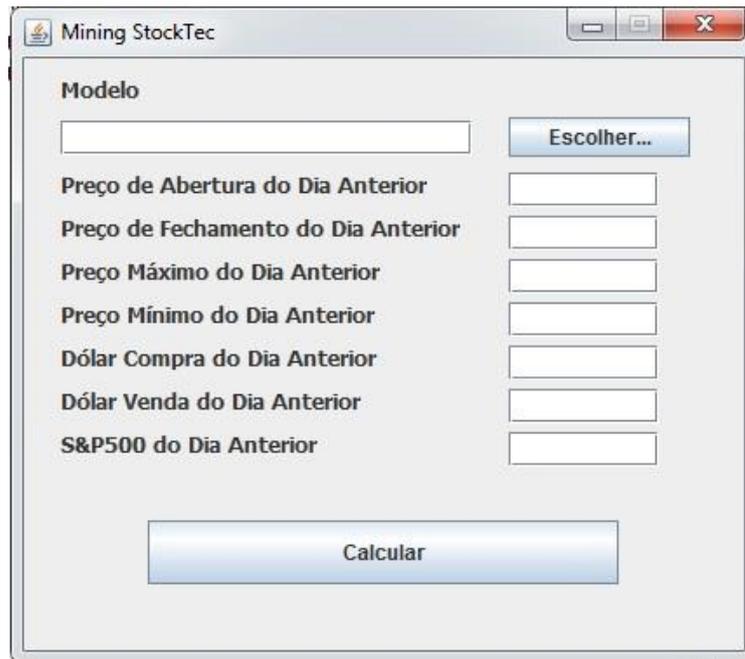


Figura 25 - Tela de Predição

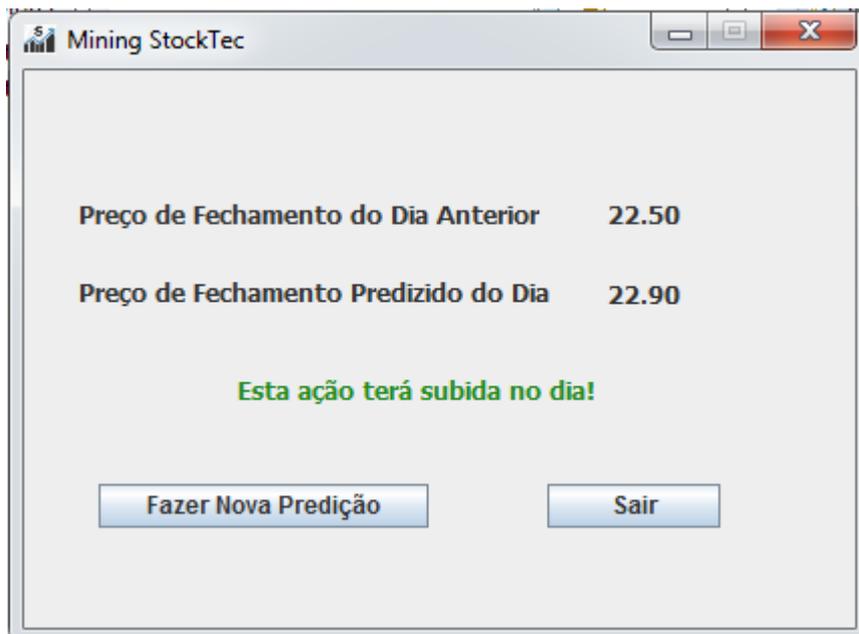


Figura 26 - Tela de Resultado do Mining StockTec

4.5.2 Mining StockTec na estratégia de investimento

Neste exemplo, espera-se utilizar o Mining StockTec para verificar se a tendência do papel VALE3 (VALE:AS) no dia 12 de maio é de alta ou baixa além de saber qual o seu possível valor de fechamento. Os valores de atributos iniciais que serão utilizados são respectivos ao dia 9 de maio, são eles:

- Valor de Abertura: 29.52
- Valor Máximo : 29.67
- Valor Mínimo: 29.11
- Valor de fechamento: 29.22
- S&P 500: 187.96
- Dólar compra e venda: 2,2150 - 2,2154

Como visto anteriormente, o Mining StockTec contém em sua base de dados um modelo gerado pelo Weka que contém a fórmula geradora do *output* das previsões (ver 4.3.2). Ao aplicar os valores e clicar em 'Calcular' (ver figura 27) o produto final será gerado e o investidor receberá a informação que deseja visualizar.

De acordo com o exemplo em questão, o Mining StockTec resulta em um valor de fechamento para o dia 12 de maio de 30.13 reais e tendência de alta.

Note que todo este processamento, é fruto de diversos passos que foram desenvolvidos de forma iterativa com o intuito de aplicar um novo conhecimento. O investidor ao visualizar tal informação pode montar uma estratégia para se posicionar nesta ação e gerar um lucro na operação. Segue, então, a estratégia:

Estratégia de posicionamento em ação:

No início do dia, tendo em posse que a perspectiva do mercado para a ação da VALE é boa e que o valor aproximado de fechamento é 30.13, o investidor pode montar uma estratégia que visa gerar lucro de 2%.

Ou seja, se o investidor comprar por 2 lotes de 100 ações por R\$ 29.90 cada, gastará: R\$ 5.980. Aplicando 2% em cima deste valor, tem-se o preço alvo da ação

em R\$ 30,50. Este é o valor da ação que investidor precisará vender para realizar um lucro de 2% na operação e zerar sua posição.

Neste dia, o mercado se comportou como o previsto pelo Mining StockTec e as 15 horas e 22 minutos a ação atingiu o valor de 30,51 centavos.

Desta forma, baseando-se na acurácia do Mining StockTec, o investidor conseguiu aplicar com êxito sua estratégia nesta operação, gerando para si um lucro de 2% referentes a R\$ 119,60 em poucas horas de negociação.

Capítulo 5 – CONCLUSÕES E TRABALHOS FUTUROS

Tendo seguido todos os passos e técnicas abordados anteriormente neste trabalho, o *output* do processo consiste em um novo conhecimento que foi inferido através do processo de KDD. Este capítulo reserva-se a analisar o produto final e alvitar temas para trabalhos futuros.

5.1 Conhecimento aprendido após aplicação

O ambiente de mercado financeiro é bastante mutável e apresenta diversas tendências ao longo de um determinado período de tempo. O uso do algoritmo de SVM no processo de descoberta de conhecimento em banco de dados, aliado a uma análise de sentimento que busca compreender qual é o “humor” do mercado através das notícias mostrou-se extremamente relevante para a operação na bolsa de valores desde que as informações obtidas sejam utilizadas de forma coerente.

Ao analisar o resultado do capítulo 4, é possível verificar que a análise de sentimento aliada à mineração de dados proposta neste trabalho foi uma associação bem sucedida.

Este trabalho demonstrou que é possível unir a mineração de dados com a análise de notícias para prever o valor final de uma ação em negociação no mercado de ações brasileiro. Continuar evoluindo a lógica utilizada no momento da aplicação da análise de sentimento, como tratamento de várias fontes de informações e relacionamento da quantidade de notícias positivas e negativas é algo a ser realizado com o intuito de refinar ainda mais os resultados da influência das notícias nas ações.

Foi proposto também, um software denominado Mining StockTec. Utilizando o modelo de dados aprendido na fase de KDD, o Mining StockTec demonstrou de forma prática o meio pelo qual o algoritmo conseguiu prever a tendência de diversos dias de negociação, tornando-se um valioso aliado do investidor no momento da tomada de decisão.

5.2 Trabalhos Futuros

Nos últimos anos as técnicas de prever o mercado e antecipar-se a suas mudanças vem ganhando força e hoje são ferramentas importantes para construir uma estratégia de investimento.

Porém, a emoção desmedida aplicada à estratégia torna-se muita das vezes um empecilho para a atividade de investimento (ver 2.1.2). Com isso, surgiram plataformas automatizadas de operação no mercado, que são programas criados através da linguagem Mlq5.

Com o Mlq5 o programador pode criar um Expert Adviser, também conhecido como Forex Robot. Este código é capaz de realizar operações financeiras baseadas em informações de mercados, *trades* passados, algoritmos e funções específicas. Grandes empresas utilizam forex robots em um servidor de alta performance para realizar trades que duram segundos através de plataformas home broker.

Como esta operação geralmente utiliza uma janela de tempo curta, é possível acompanhar pequenos movimentos no mercado com uma operação de compra e venda de ações. Após ser repetida diversas vezes de forma automática, a operação dos robôs de investimento resultam em um considerável lucro ao fim do dia de negociação.

Trabalhos futuros incluem apresentar uma outra forma de introduzir as notícias, avaliação para diversas empresas em tempo real com o mercado ou uma análise analítica da linguagem Mlq5, além da aplicação do processo de descoberta de conhecimento para melhorar a operação automatizada de compra e venda de papéis acionários. Estendendo o conceito anterior, pode-se incluir a análise de sentimento do que é discutido em tempo real através das redes sociais e aliar ao processo de criação de um *forex robot* mais inteligente e mutável ao ambiente de mercado no qual se encontra inserido.

Referências Bibliográficas

BEN-DAVID, S.; LINDENBAUM, M. (1997). Learning distributions by their density levels: A paradigm for learning without a teacher, *Journal of Computer and System Sciences* 55, 171-182.

LIU, Bing (2008). Mining Opinions in Comparative Sentences. In: 22nd International Conference on Computational Linguistics (Coling-2008), Manchester.

CAO, L.J.; TAY, F. (2001). Application of support vector machines in financial time series forecasting. [S.l]: *International Journal Of Management Science*.

CORTES, C.; VAPNIK, V. N (1995). Support vector networks: *Machine Learning* 20. [S.l]: Springer.

FAYYAD, Usama; PIATETSKI-SHAPIRO, Gregory; SMYTH, Padhraic (1996) The KDD Process for Extracting Useful Knowledge from Volumes of Data. In: *Communications of the ACM*, pp.27-34.

HAN, Jiawei; KAMBER, Micheline (2011). *Data Mining: Concepts and Techniques*. 3. ed.[S.l]: Morgan Kaufmann.

JACKSON, Peter; MOULINIER, Isabelle. John benjamins publishing company (2007). [S.l]: Cambridge University Press.

JOACHIMS, T. (2002). *Learning to Classify Text using Support Vector Machines*, Kluwer Academic Publishers, London.

KLOSGEN, W. (1992) Patterns for Knowledge Discovery in Databases. *Proc. Of Machine Learning*. UK, p. 1-9.

LIAO, T. W.; CHEN, J. H.; TRIANTAPHYLLOU, E. (1999). Data mining applications in industrial engineering: A Perspective. In: 25 th International Conference on Computers and Industrial Engineering, New Orleans, LA.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. Introduction to information retrieval (2008). [S.l]: Cambridge University Press.

MITCHELL, Tom M.. Machine Learning (1997). [s.l]: Mcgraw-hill Science/engineering/math.

MOORE, Marcos (2012). Ações: Quais comprar e quando comprar: aprenda a investir utilizando análise fundamentalista com análise gráfica.[S.l]:Campus.

MUKHERJEE, S., OSUNA, E., & Girosi, F. (1997). Nonlinear prediction of chaotic time series using support vector machine. Proceedings of the IEEE workshop on Neural Networks for Signal Processing, Amelia Island, FL., pp. 511-520.

OSUNA, E.E, FREUND, R. (1997). Support Vector Machines: Training and Applications, M press, USA.

OU, Phichhang; WANG, Hengshan (2009). Prediction of stock market index movement by ten data mining techniques. Shangai: Canadian Center of Science and Education.

PRESS, William H (2007). et al. Numerical recipes 3rd edition: the art of scientific computing . 3. ed. [S.i]: Cambridge University Press.

TARASSENKO, L. (1995) et al. Novelty detection for the identification of masses in mammograms. In: Proceedings fourth IEE International Conference on Artificial Neural Networks, Cambridge.Pages 442-447.

VAPNIK, V.N (1995). et al. The nature of statistical learning theory, New York, Springer-Verlag.

WOLWACZ, Alexandre (2007). Táticas Operacionais de Posição em Ações. 3. ed.[S.l]: Leandro & Stormer.