UNIVERSIDADE FEDERAL DO MARANHÃO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
ÁREA DE CIÊNCIA DA COMPUTAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA
DE ELETRICIDADE

Um Modelo para Predição de Bolsa de Valores Baseado em Mineração de Opinião

Milson Louseiro Lima

Dissertação de Mestrado

São Luís-MA
06 de maio de 2016

Milson Louseiro Lima

Um Modelo para Predição de Bolsa de Valores Baseado em Mineração de Opinião

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Eletricidade, da Universidade Federal do Maranhão, como requisito para o título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Sofiane Labidi

São Luís-MA
06 de maio de 2016

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Núcleo Integrado de Bibliotecas/UFMA

Lima, Milson.
Um Modelo para Predição de Bolsa de Valores Baseado em Mineração de Opinião / Milson Lima. - 2016.
113 p.

Orientador(a): Sofiane Labidi.
Dissertação (Mestrado) - Programa de Pós-graduação em Engenharia de Eletricidade/ccet, Universidade Federal do Maranhão, São Luis-MA, 2016.

1. Análise de Sentimento. 2. Bolsa de Valores. 3. Inteligência Artificial. 4. Mineração de Opinião. 5. Twitter. I. Labidi, Sofiane. II. Título.

UM MODELO PARA PREDIÇÃO DE BOLSA DE VALORES BASEADO EM MINERAÇÃO DE OPINIÃO

MILSON LOUSEIRO LIMA

Prof. Sofiane Labidi, Dr.
(Orientador)

Prof.ª Karla Donato Fook, Dr.ª
(Membro da Banca Examinadora)

Prof. Denivaldo Cicero Pavao Lopes, Dr.
(Membro da Banca Examinadora)

‘‘A utopia está no horizonte. Aproximo-me dois passos, ela se afasta dois passos.
Caminho dez passos e o horizonte se distancia dez passos mais além.
Para que serve a utopia? Serve para isso: para caminhar.’’

Eduardo Galeano

Agradecimentos

A Deus, por ter me concedido a oportunidade de ter chegado até aqui. A meus familiares, a meu pai (*in memorian*) e em especial a minha mãe que tanto lutou para me dar o direito à educação.

Ao meu orientador Prof. Dr. Sofiane Labidi, pela oportunidade proporcionada em integrar a equipe de pesquisadores do Laboratório de Sistemas Inteligentes. Sou grato pela paciência e dedicação dispensada durante todo esse período.

Aos professores Zair Abdelouahab, Maria Del Rosário Girardi e Nilson Santos Costa, pelos ensinamentos transmitidos que foram de grande valia em todos os aspectos, sejam acadêmicos, sejam na vida pessoal.

Aos amigos e companheiros desta Universidade, Christian Diniz, Thiago Pinheiro, Nádson Timbó, Pedro Brandão, Guilherme Lima, Gilberto Nunes, Rafael Pinheiro, Alexandre Ataíde e Isaac Júnior, pelo apoio e pela parceria durante essa jornada.

A Filipe Lima, por ser a inspiração diária na busca de um futuro melhor, sendo a razão de todos os sacrifícios.

Por fim, a todos que contribuíram de forma direta ou indireta para a realização deste trabalho de pesquisa e conclusão do curso.

Resumo

Predizer o comportamento das ações na bolsa de valores é uma tarefa desafiadora, muita vezes relacionada a fatores desconhecidos ou influenciados por variáveis de naturezas bem distintas, que podem ir desde notícias de grande repercussão até o sentimento coletivo, expresso em publicações de redes sociais. Tal volatilidade do mercado pode representar perdas financeiras consideráveis para os investidores. No intuito de se antecipar a tais variações já foram propostos outros mecanismos para predizer o comportamento de ativos na bolsa de valores, baseados em dados de indicadores pré-existentes. Tais mecanismos analisam apenas dados estatísticos, não considerando o sentimento humano coletivo. Este trabalho tem como finalidade desenvolver um modelo para predição da bolsa de valores, baseado na mineração de opinião e, para isso, fará uso de técnicas de Inteligência artificial como processamento de linguagem natural(PLN) e Máquinas de Vetor de Suporte(SVM) para predizer o comportamento do ativo. No entanto, convém ressaltar que o referido modelo tem como finalidade ser uma ferramenta de auxílio no processo de tomada de decisão que envolve a compra e venda de ações na bolsa de valores.

Palavras-chave: *Mineração de Opinião; Análise de Sentimento; Twitter; Bolsa de Valores; Inteligência Artificial.*

Abstract

Predicting the behavior of stocks in the stock market is a challenging task, a lot of times related to unknown factors or influenced by very distinct natures of variables, which can range from high-profile news to the collective sentiment, expressed in publications on social networks. Such market volatility may represent considerable financial losses for investors. In order to forestall such variations other mechanisms to predict the behavior of assets in the stock market have been proposed, based on pre-existing indicator data. Such mechanisms only analyze statistical data, not considering the collective human sentiment. This work aims to develop a model to predict the stock market, based on analysis of sentiment and it will make use of techniques of artificial intelligence as natural language processing (PLN) and Support Vector Machines (SVM) to predict the active behavior. However, it should be emphasized that this model is intended to be an aid tool in the decision-making process that involves buying and selling shares on the stock market.

Keywords: *Opinion mining; Sentiment Analysis; Twitter; Stock Market; Artificial intelligence.*

Lista de Figuras

Lista de Tabelas

Lista de Siglas e Abreviaturas

BOVESPA	Bolsa de Valores de São Paulo
TIC	Tecnologias de Informação e Comunicação
PLN	Processamento de Linguagem Natural (Natural Language Processing)
API	Interface de Programação de Aplicativos (Application Programming Interface)
KDD	Descoberta de Conhecimento em Base de Dados (Knowledge Discovery in Databases)
SVM	Máquinas de Vetor de Suporte (Support Vector Machines)
HSX	Bolsa de Valores de Hollywood (Hollywood Stock Exchange)
DJIA	Dow Jones Industrial Average
LSI	Laboratório de Sistemas Inteligentes
WEKA	Waikato Environment for Knowledge Analysis
UML	Unified Modeling Language

Sumário

Capítulo 1

Introdução

O presente trabalho contextualiza a difícil tarefa de prever o comportamento do mercado financeiro e propõe um modelo para predição da bolsa de valores, baseado na mineração de opinião a partir de mensagens do *Twitter*[1], considerando o sentimento coletivo resultante dessas mensagens, uma das variáveis determinantes para definir o comportamento financeiro de um ativo na bolsa de valores.

1.1 Contexto e Motivação

O desenvolvimento econômico e sustentável de um país depende da expansão contínua de sua capacidade de produção. Esta relação é função, por sua vez, de investimentos em capital e recursos humanos. O crescimento se acelera quando os investimentos se direcionam para as alternativas com maiores retornos econômicos e sociais Bovespa (2000).

Uma das principais formas de crescimento econômico ocorre através do mercado de capitais, pois o mesmo constitui-se de mecanismos eficientes que incentivam a formação de poupança, possibilitam sua intermediação e são de fácil acesso a quem deseja investir. Esse mercado, constituído pelas bolsas de valores e corretoras, é bastante eficiente para captar poupança e canalizá-la para as atividades mais produtivas BMFBovespa (2010).

Nesse contexto, predizer o comportamento da bolsa de valores é um diferencial estratégico que pode representar ganhos significativos. Profissionais da área se dedicam a estudar diversos tipos de indicadores que auxiliem no processo de tomada de decisão e redução das incertezas quanto ao mercado.

[1]Disponível em: http://twitter.com

Este trabalho propõe desenvolver uma abordagem utilizando Mineração de Opinião (Análise de Sentimentos) no *Twitter*, a fim de mensurar a polaridade do que é expresso pelos usuários, objetivando estabelecer uma relação entre o sentimento coletivo e o comportamento financeiro de determinada empresa no Mercado Financeiro.

1.2 Problemática

A bolsa de valores exerce um papel de extrema importância para o desenvolvimento econômico de uma sociedade, no entanto, predizer seu comportamento ainda é algo desafiador, que envolve grandes riscos ao capital.

Ao longo das últimas décadas, a utilização de modelos estatísticos tradicionais para prever o comportamento da bolsa de valores vem ganhando auxílio de outras abordagens, principalmente as que utilizam inteligência artificial e suas técnicas.

Segundo Graeml (1998), existem métodos tradicionais que auxiliam na tomada de decisão para o mercado financeiro, contudo, tais métodos já não apresentam a mesma eficiência. Desta forma, há uma grande necessidade de novas abordagens preditivas para a bolsa de valores.

Torna-se oportuno salientar que investir na bolsa de valores é uma atividade que envolve riscos, pois o mercado financeiro é repleto de incertezas, podendo reagir de forma inesperada a determinados eventos externos.

Conforme Chorafas (1992), um evento de risco é um episódio incerto, que não apresenta resultado definido. Ainda de acordo com o autor, os riscos estão relacionados a uma situação, posição ou escolha que envolve possíveis perdas, devido à incerteza dos seus resultados, de modo que o risco seria o custo da incerteza.

Sendo assim, dispor de mecanismos que auxiliam na tomada de decisão e que venham a minimizar a exposição aos riscos, inerentes ao mercado, contribuindo para a maximização dos lucros, revela-se uma ferramenta de grande importância.

Nesse cenário de busca por novas abordagens para predição de ativos da bolsa de valores, as redes sociais surgem como um ambiente em expansão, onde o compartilhamento de informações e a grande quantidade de dados gerados pela interação entre os usuários têm alcançado ampla notoriedade nos últimos anos, tornando-se um local atrativo para pesquisas em diversas áreas.

Para Zhang et al. (2010), as redes sociais têm sido alvo de diversos estudos, particularmente em termos de previsão e descrição do mercado acionista.

A utilização de meios preditivos baseados nas técnicas de Inteligência artificial, vem ganhando grande destaque, sendo que uma das mais recentes abordagens diz respeito à mineração de opinião. Este trabalho fará uso deste conceito para predizer o comportamento da bolsa de valores, partindo da análise dos sentimentos expressos nas mensagens do *Twitter*.

1.3 Objetivos

O presente trabalho tem como principal objetivo desenvolver e apresentar um modelo para predição na bolsa de valores, que a partir da análise dos sentimentos, extraídos de mensagens do *Twitter*, seja capaz de prever a tendência de fechamento de uma ação de determinada empresa no mercado financeiro, estabelecendo, assim, uma relação entre o sentimento coletivo e o comportamento do ativo em dado período.

Tendo em vista alcançar o objetivo principal, torna-se necessário cumprir as seguintes etapas representadas pelos objetivos específicos:

- Fazer o levantamento bibliográfico sobre as áreas envolvidas na pesquisa;
- Analisar os trabalhos relacionados existentes;
- Elaborar uma análise sobre as técnicas de predição no contexto da bolsa de valores, bem como suas principais limitações;
- Desenvolver técnicas para a coleta e preparação de *dataset's*[2] a partir da fonte de dados escolhida (*Twitter*).
- Aprofundar os estudos sobre o Processo de Descoberta de Conhecimento (KDD) e suas diversas etapas, especialmente a mineração de dados;
- Selecionar as técnicas, ferramentas e algoritmos para mensurar o sentimento social;
- Processar as informações utilizando algoritmo de aprendizagem de máquina;
- Validar o modelo proposto.

[2] É um subconjunto de dados resultante de uma consulta a um banco de dados.

1.4 Metodologia da Pesquisa

A presente pesquisa faz uso em seus estudos de técnicas de Mineração de Dados em Redes Sociais, bem como algoritmo de aprendizagem de máquina, tendo por objetivo, predizer o comportamento de fechamento do mercado de ações para um determinado ativo na bolsa de valores.

Para esta pesquisa utilizou-se o método hipotético-dedutivo, proposto por Popper (1980), que consiste em solucionar um problema através de tentativas (conjunturas, hipóteses, teorias) e eliminação de erros. A baixo demais características sobre a metodologia aplicada na pesquisa:

- Levantamento bibliográfico (vide objetivos específicos);
- Tipo da pesquisa: Exploratória;
- Universo da pesquisa: A rede social *Twitter*;
- Amostragem: Dados coletados do *Twitter*, originários de publicações sobre a empresa "Petrobras" no período compreendido entre setembro e novembro do ano de 2015.
- Coleta dos dados: Os dados foram coletados utilizando-se dos recursos disponibilizados pela API do *Twitter* em conjunto com aplicação desenvolvida para esta finalidade que será melhor detalhada no Cap.4 na seção 4.2;
- Resultados: A pesquisa propõe um estudo de caso onde avalia a opinião coletada no *Twitter* e seu relacionamento com o mercado financeiro, estabelecendo assim, um índice de acertos (acurácia) entre os dados reais em uma série histórica de fechamento da bolsa;
- Modelo: Ao fim da aplicação da metodologia, tem-se como resultado um modelo que visa ser uma ferramenta de auxilio ao processo de tomada de decisão de compra e venda de ativos.

1.5 Justificativa e Relevância

A bolsa de valores é uma oportunidade que o investidor tem à sua disposição para formar um patrimônio e, ao mesmo tempo, fornecer recursos para o crescimento das organizações Bovespa (2000).

Segundo Souza (2005), a economia cresce mediante estímulos específicos, que se transmitem, ao sistema como um todo, transformando-o. Nesse norte, pode-se concluir que o investimento é um dos fatores essenciais para promover o crescimento econômico.

Diante das oportunidades financeiras que o ato de investir na bolsa de valores pode representar, existem vários bons motivos para se canalizar esforços na tentativa de predizer o Mercado Financeiro. Um dos principais é a vantagem que uma ferramenta de predição com grandes índices de acertos representaria para seu detentor ou manipulador frente aos demais negociadores. A capacidade de poder se antecipar ao movimento de um mercado com tantas variáveis que influenciam na oscilação dos preços dos ativos, pode representar grande lucratividade.

Um fenômeno crescente nos dias atuais é a preocupação das empresas com sua imagem, figurando as redes sociais, nesse contexto, como o ambiente propício para que seus usuários, sejam eles clientes ou não, manifestem suas opiniões a respeito de produtos, serviços ou sobre algum aspecto que julguem pertinentes. O sentimento social contido nessas publicações pode exercer um papel importante na formação de opinião sobre suas marcas, o que pode vir a gerar reflexos sob o ponto de vista econômico, repercutindo no valor dos seus ativos na bolsa de valores.

Neste cenário, cada vez mais pesquisadores, das mais diferentes áreas, dedicam-se em construir modelos no intuito de desenvolver ferramentas que sejam capazes de predizer o mercado financeiro, obtendo os melhores resultados possíveis. Nessa busca, recentemente foi inserida a utilização da Inteligência Artificial, Aprendizagem de Máquina e Mineração de Dados, ganhando bastante espaço entre os pesquisadores. Acerca do tema, pode-se citar o trabalho de Bollen et al. (2011) que procurou relacionar os *Twitter feeds* ao valor do *Dow Jones Industrial Average* (DJIA) ao longo de um determinado período, baseando-se no humor (calma, alerta, Claro, Vital, bondoso e feliz) expresso pelos usuários. Outro trabalho que merece destaque é o de Marques (2010), que desenvolveu um sistema inteligente, utilizando o indicador de análise técnica do mercado financeiro MACD com o emprego de algoritmos genéticos e lógica *fuzzy*.

Diante da importância do mercado financeiro para uma economia e da popularização do uso das redes sociais, torna-se de grande valia dispor de uma ferramenta que possa fazer uso desses dados, no intuito de criar um indicador de auxílio no complexo processo de tomada de decisão que envolve a bolsa de valores.

Este trabalho, através da avaliação dos sentimentos contidos nas mensagens do *Twitter*,

fará uma abordagem complementar no que diz respeito às atuais ferramentas de auxílio usadas na tomada de decisão que tem como finalidade a predição da bolsa de valores. Neste modelo, as postagens funcionam como uma espécie de termômetro social, indicando o humor coletivo dos usuários em relação a uma determinada empresa no mercado financeiro, de maneira a tornar essa informação util quando comparada com os indicadores do mercado, almejando auxiliar o investidor a maximizar sua rentabilidade e minimizar suas incertezas no momento de comprar ou vender um ativo.

1.6 Contribuições Científicas

A predição do mercado financeiro é um assunto que vem sendo estudado por diversas áreas, fazendo uso de técnicas, como os métodos estatísticos lineares, baseados em indicadores financeiros, e, mais recentemente, abordagens não lineares vêm ganhando bastante notoriedade devido seus resultados promissores. Neste contexto, o presente trabalho, utilizando-se de técnicas de Inteligência Artificial, visa demonstrar que a inserção da variável "sentimento", no contexto de predição que utiliza indicadores financeiros, proporciona um ganho significativo nos resultados. Disso advém uma abordagem complementar no amparo à tomada de decisão que envolve os ativos de uma empresa no mercado de ações.

1.7 Estrutura do Trabalho

Além deste capítulo introdutório, o trabalho em tela conta com mais 5 (cinco) capítulos, os quais estão estruturados da seguinte forma:

- Capítulo 2 - Fundamentação Teórica: será apresentado o referencial teórico, onde são abordados aspectos teóricos que serviram de base para a pesquisa;
- Capítulo 3 - Trabalhos Relacionados: serão mostrados trabalhos relacionados ao tema que contribuíram para o desenvolvimento desta pesquisa;
- Capítulo 4 - Um Modelo para Predição de Bolsa de Valores Baseado em Mineração de Opinião: neste capítulo são demonstrados os principais aspectos referentes à implementação, testes, prototipação e aplicação da abordagem proposta;
- Capítulo 5 - Resultados Obtidos: aqui serão mostrados os resultados obtidos com a pesquisa;

- Capítulo 6 - Conclusões: serão mostrados a conclusão da pesquisa, artigos publicados ou aceitos e sugestões para trabalhos futuros, que venham a melhorar os resultados deste estudo.

Capítulo 2

Fundamentação Teórica

Este capítulo destina-se a mostrar conceitos relevantes para a fundamentação e a construção da proposta do trabalho, assim como contribuir para um melhor entendimento da metodologia aplicada. Aborda-se, dentre outros assuntos, o mercado financeiro brasileiro, a predição da bolsa de valores, as redes sociais e, por fim, a descoberta do conhecimento e a mineração de opinião.

2.1 O Mercado Financeiro e a Predição da Bolsa de Valores

O Mercado Financeiro é o local onde as pessoas negociam o dinheiro, mesmo estando expostas aos riscos inerentes ao mercado. Responsável por estabelecer a negociação entre os agentes econômicos[1] que têm um excedente de capital e os que precisam de dinheiro, o Mercado Financeiro, sob o intermédio de um agente financeiro, promove essa transação entre superavitários e deficitários, como mostra a Figura 2.1.

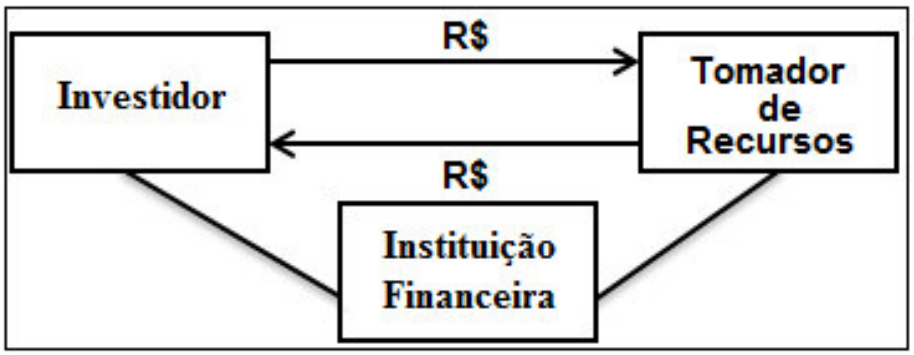


Figura 2.1: Representação da Interação Entre os Agentes no Mercado Financeiros (Andreso e Lima, 2007, p.3).

[1] Agentes incluem por exemplo, indivíduos, governo, empresas ou organizações

A mencionada negociação entre os agentes no mercado financeiro é de extrema importância para o desenvolvimento de uma economia, pois permite a elevação das taxas de poupança e investimento, sendo responsável por englobar qualquer operação financeira nas suas mais diversas formas, sejam moedas, debêntures, ações, dentre outras representações do dinheiro.

Para Fortuna (2008), o sistema financeiro é uma conceituação bastante abrangente, sendo assim, um conjunto de instituições financeiras que se dedicam, de alguma forma, ao trabalho de proporcionar aos poupadores e investidores condições que satisfaçam a manutenção de um fluxo de recursos.

Interessante pontuar que Sandroni (1994) define o Mercado Financeiro como sendo um conjunto formado pelo mercado monetário e pelo mercado de capitais, abrangendo todas as transações com moedas e títulos e as instituições que as promovem: banco central, caixas econômicas, bancos estaduais, bancos comerciais e de investimentos, corretoras de valores, distribuidoras de títulos, fundos de investimento, bolsas de valores etc. Já Santos (2000), de forma bem sucinta, mas não menos importante, menciona que o mercado nada mais é do que um grande fundo, do qual pode-se depositar e sacar de acordo com uma determinada taxa de juros.

Segundo Andreso e Lima (2007), o mercado financeiro, em regra, está segmentado do seguinte modo:

- **Mercado Monetário**: composto pelo conjunto de operações de curto e curtíssimo prazo. Tais operações se destinam a atender às necessidades imediatas de liquidez dos agentes econômicos;

- **Mercado de Crédito**: composto pelo conjunto de operações de prazo curto, médio ou aleatório. Destina-se basicamente a suprir as necessidades de caixa de curto e médio prazo de indivíduos ou empresas;

- **Mercado Cambial**: É composto pelo conjunto de operações de compra e venda de moedas de diferentes países;

- **Mercado de Capitais**: É composto pelo conjunto de operações de prazo médio, longo ou indeterminado. Destina-se, principalmente, ao financiamento de capital fixo. Esse mercado abrange, por exemplo, debêntures, *bonds* e *notes*, e, no caso de operações de prazo indeterminado, as ações. Para Assaf Neto (2006), o referido mercado tem papel

relevante no desenvolvimento econômico, visto que é uma grande fonte de recursos para investimentos da economia.

Andreso e Lima (2007) ressaltam que a segmentação acima é meramente didática, visto que na prática tal divisão se confunde em decorrência da vastidão e da complexidade das operações realizadas no mercado. A Figura 2.2 ilustra o citado fracionamento.

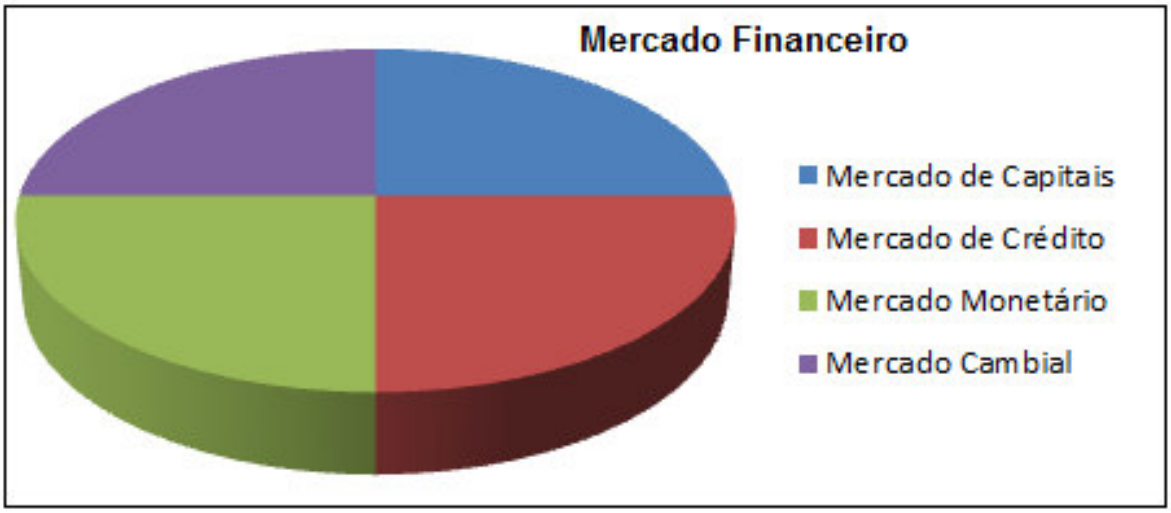


Figura 2.2: Segmentação do Mercado Financeiro Nacional. Adaptado de Lopes (2012)

Na próxima subseção serão demonstrados conceitos referentes à bolsa de valores, os quais ajudarão a entender a dinâmica desse mercado.

2.1.1 Bolsa de Valores

São locais que oferecem condições e sistemas necessários para a realização de negociação de compra e venda de títulos e valores mobiliários de forma transparente, possuindo atividade de auto regulação que visa preservar elevados padrões éticos de transação e divulgar as operações executadas com rapidez, amplitude e detalhes BMFBovespa (2010).

De acordo com Pinheiro (2008), existem três figuras que participam ativamente do mercado de negociação da bolsa de valores: os especuladores [2], que não se preocupam com as ações que estão comprando, mas que visam lucrar explorando a liquidez e a volatilidade do mercado

[2] É o indivíduo que age no mercado de ações visando o lucro no curto ou médio prazos.

de ações; os gestores financeiros [3], que gerenciam empresas, captando recursos a baixo custo e realizando investimentos sem riscos; e os investidores [4], os quais se utilizam do mercado visando lucro a longo prazo, através das operações de compra e venda de títulos.

Conforme Luquet e Rocco (2005), as ações podem ser negociadas nas bolsas das seguintes formas:

- **Mega Bolsa**: sistema de negociação que engloba o pregão viva-voz e os terminais remotos;
- ***Home Broker***: nesse sistema a corretora pode cumprir as ordens de seus clientes no seu próprio escritório, por meio da internet. Os computadores registram as ofertas de compra e venda e quando estas são efetivadas o negócio é realizado eletronicamente;
- ***After-market***: ampliação do horário do pregão eletrônico, que funciona das 18h00min às 22h00min para atender o investidor que opera via internet.

2.1.2 Ações

A ação é um título de renda variável que representa a menor fração do capital da empresa emitente. Quando um investidor compra uma ação, torna-se coproprietário desta empresa, participando de seus resultados BMFBovespa (2010).

Na BOVESPA, são negociadas ações de empresas brasileiras (mercado a vista), bem como seus derivativos(mercados a termo, de opções e futuro de ações) e certificados de ações de empresas estrangeiras.

Ações são títulos nominativos negociáveis que representam uma fração do capital social de uma determinada empresa. No Brasil, a Lei das Sociedades por Ações [5] permite que as empresas emitam dois tipos de ações, a dizer:

- **Ordinárias (On)**: asseguram o direito de voto nas assembleias de acionistas da empresa;
- **Preferenciais (Pn)**: oferecem preferência no recebimento de resultados ou no reembolso do capital em caso de liquidação da companhia. Entretanto, as ações preferenciais não concedem o direito de voto ou o restringem.

[3] É o profissional responsável por coordenar a tesouraria e/ou controladoria de uma empresa ou instituição.
[4] São indivíduos ou instituições que aplicam recursos em busca de ganhos a médio e longo prazos e que operam nas Bolsas por meio de Corretoras e Distribuidoras de Valores.
[5] É a lei que rege as Sociedades Anônimas (capital da empresa não está associado a um único nome). Data de 15 de dezembro de 1976. Disponível em: http://www.planalto.gov.br

Os investimentos em ações permitem aos seus respectivos possuidores diversos direitos e obrigações, esta última unicamente o fato de que, uma vez que tenha subscrito ações de um aumento de capital, torna-se obrigado a integralizar, ou seja, pagar o valor das ações que subscreveu Pinheiro (2008).

Para identificação na bolsa de valores, as ações e demais ativos negociados recebem um código de quatro letras e um número. As letras indicam o nome da empresa e o número o tipo de ação. Na Tabela 2.1 é possível vizualizar um exemplo de cotação no *homebroker* e suas informações.

Tabela 2.1: Modelo de Tabela de Cotação do *homebroker*.

COD	**Desc.Ativo**	**VAR**	**ULT**	**OC**	**OV**	**MAX**	**MIN**	**ABE**	**FEC**	**VOL**
CSNA3	Sid. Nacional ON	-1.2%	13.13	13.12	13.13	13.5	12.96	13.29	13.29	5.775.200
ELET6	Eletrobras PNB	-1.1%	20.8	20.7	20.8	21.31	20.51	21.19	21.03	1.191.400
ETER3	Eternit ON	3.1%	10.79	10.66	10.79	10.93	10.34	10.4	10.47	248.700
GGBBR4	Gerdal PN	0.5%	15.57	15.56	15.57	15.73	15.21	15.55	15.5	10.867.00
INEP4	Inepar PN	5.0%	2.09	2.06	2.09	2.09	1.93	2.0	1.99	226.900
JHSF3	Jhsf Part ON	-1.7%	5.85	5.85	5.88	6.01	5.8	5.95	5.95	766.200
MILK11	Laep DR3	0.0%	0.15	0.15	0.16	0.16	0.14	0.15	0.15	5.386.300
PETR3	Petrobras ON	3.0%	19.78	19.77	19.8	19.99	19.23	19.4	19.2	8.937.300
PETR4	Petrobras PN	3.5%	19.07	19.07	19.09	19.21	18.43	18.6	18.43	30.827.600
PETR3	Petrobras ON	3.0%	19.78	19.77	19.8	19.99	19.23	19.4	19.2	8.937.300
PETR4	Petrobras PN	3.5%	19.07	19.07	19.09	19.21	18.43	18.6	18.43	30.827.600

Fonte: http://mercadoreal.net

Onde:

- **COD**: Código de negociação do ativo.
- Descrição do Ativo: Nome da empresa e ativo de ação (Cadastro Bovespa).
- **VAR(%)**: Variação percentual do preço do ativo em relação ao preço de fechamento do pregão anterior.
- **ULT** – Último Negócio: Valor do último negócio realizado no ativo
- **OC** – Oferta de Compra: Maior preço ofertado para a compra do ativo no momento.
- **OV** – Oferta Venda: Menor preço ofertado para a venda do ativo no momento.
- **MAX** – Preço Máximo do dia: É o preço mais alto que o ativo foi negociado.
- **MIN** – Preço Mínimo do dia: É o preço mais baixo que o ativo foi negociado.
- **ABE** – Preço de Abertura: É o preço do primeiro negócio realizado no pregão.

- **FEC** – Preço de Fechamento: É o preço do último negócio realizado no dia.
- **VOL** – Volume: É a quantidade de títulos negociados no pregão.

2.1.3 Métodos para Análise do Mercado e Predição da Bolsa de Valores

A capacidade de prever o futuro apuradamente é fundamental para muitos processos de decisões em planejamento, programação, aquisição, formulação da estratégia, definição de políticas e operações de cadeia de suprimentos Zhang (2004). No mercado financeiro não é diferente, a utilização de boas ferramentas que auxiliem a tomada de decisão é indispensável e pode representar o sucesso ou o fracasso em uma negociação. Em seguida serão explanados métodos para esta finalidade. Existem no mercado de capitais dois modelos de avaliação que procuram compreender e propor estratégias para a compra e venda de ações de uma determinada empresa: o modelo fundamentalista e o modelo técnico Limeira (2003).

2.1.3.1 Fundamentalista

A análise fundamentalista busca avaliar a situação financeira das empresas através de aspectos contábeis, macro e micro econômicos, visando projetar seus resultados futuros. Segundo Thomsett (1998), a referida análise também pode ser definida como uma metodologia de estudo de informações financeiras básicas, visando prognosticar lucros, oferta e demanda, potencial do setor em que a empresa está inserida, na habilidade gerencial e outras questões que afetam o valor de mercado das ações.

Winger e Frasca (1995) afirmam que a análise fundamentalista se sustenta com base em três fatores: análise da empresa, análise do contexto onde a empresa está inserida e análise geral da economia ou macroeconomia.

É oportuno acentuar que no modelo fundamentalista a análise conjunta das informações é de extrema importância para a tomada de decisão.

2.1.3.2 Técnica

Também conhecida como análise gráfica, baseia-se em dados históricos das ações da empresa no mercado, propondo através de ferramentas gráficas e técnicas matemáticas estabelecer padrões identificados na representação gráfica das variações que os preços demonstram em um determinado período de tempo Tavares (1987).

De uma forma geral, a análise técnica parte do princípio de que o comportamento do mercado tende a mostrar comportamentos no futuro, baseado na repetição de padrões do passado para predizer a melhor hora de comprar ou vender uma ação.

Diferentemente da análise fundamentalista, que busca mensurar os aspectos econômicos, financeiros e operacionais da empresa e suas tendências para predizer o seu comportamento, a análise técnica pauta-se em seus indicadores gráficos para determinar um padrão de crescimento histórico que possa caracterizar uma tendência nos preços dos seus ativos.

2.1.3.3 Abordagens de Predição da Bolsa de Valores

Os métodos de predição lineares utilizados durante muito tempo, tais como ARIMA (Modelo Autoregressivo Integrado de Médias Móveis) e ARMA (Modelo Autoregressivo de Médias Móveis), demonstram certa fragilidade para predizer o mercado de ações, devido a complexidade da análise e dos ruídos apresentados. Recentemente, abordagens não lineares têm sido propostas e usadas, tais como *AutoRegressive Conditional Heteroskedasticity* (ARCH), *Generalized AutoRegressive Conditional Heteroskedasticity* (GARCH), Redes Neurais Artificiais (RNAs) e Máquinas de Vetor Suporte (SVMs). Todos esses métodos têm mostrado resultados promissores Yeh et al. (2011). O presente trabalho fará uso de Inteligência Artificial, sob a forma de mineração de opinião, para predizer o comportamento do mercado. Este tipo de análise vem sendo bastante utilizada para várias finalidades e a predição da bolsa é uma delas.

2.2 Redes Sociais

O avanço e a popularização das TIC's (Tecnologias de Informação e Comunicação) disseminaram uma das mais importantes formas de expressão social da atualidade, as redes sociais.

Sinaliza-se que uma rede social representa uma estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, que partilham valores e propósitos comuns Ferreira (2011). É nesse ambiente que está o local propício para a exposição natural dos indivíduos, seus anseios, preferências e manifestações, nas suas mais diversas formas, o que torna a informação resultante desse processo mais natural e próxima da realidade.

Buscando encontrar o sentimento coletivo sobre determinado assunto, neste estudo, será explorada uma das características principais das redes sociais, que é a sua grande capacidade

de propagar as informações sob a forma de mensagens. Para isso, o *Twitter* será utilizado como fonte de dados.

2.2.1 Contextualização

As redes sociais têm adquirido importância crescente na sociedade moderna, sob vários aspectos. Sua notável capacidade de autogeração, horizontalidade e descentralização contribui para o compartilhamento de informações, conhecimentos e interesses.

Os primeiros indícios da teoria das redes encontram-se principalmente em trabalhos matemáticos, em que foi criado o primeiro teorema da teoria dos grafos. Um grafo é uma representação de um conjunto de nós conectados por arestas que, em conjunto, formam uma rede. Para Emirbayer e Goodwin (1994), uma rede social é um grafo, orientado ou não, que mapeia uma realidade ou um mundo restrito, no qual os nodos representam as entidades (indivíduos ou classes de indivíduos – também chamados atores) e as arestas representam os relacionamentos entre essas entidades. Os relacionamentos podem ser o compartilhamento de um ou mais atributos. A realidade representada pelas redes sociais é fonte de dados heterogêneos e multi-relacionais, cujos relacionamentos podem ser unidirecionais e não necessariamente precisam ser binários.

Uma das características fundamentais na definição das redes é a sua abertura e porosidade, possibilitando relacionamentos horizontais e não hierárquicos entre os participantes. As redes "não são, portanto, apenas uma outra forma de estrutura, mas quase uma não estrutura, no sentido de que parte de sua força está na habilidade de se fazer e desfazer rapidamente" Duarte e Frei (2008).

Atualmente, as redes sociais são uma das ferramentas mais atrativas disponíveis na internet. Impulsionadas, dentre outras coisas, pelo seu dinamismo inerente, elas vêm ganhando a cada dia mais adeptos, dando origem ao que pode ser chamado de uma sociedade virtual paralela, dotada de personalidade, opiniões e conceitos próprios. A Figura 2.3 apresenta um ranking das redes sociais, considerando a quantidade de usuários, dando ideia do quanto é gigantesco o universo que as envolve.

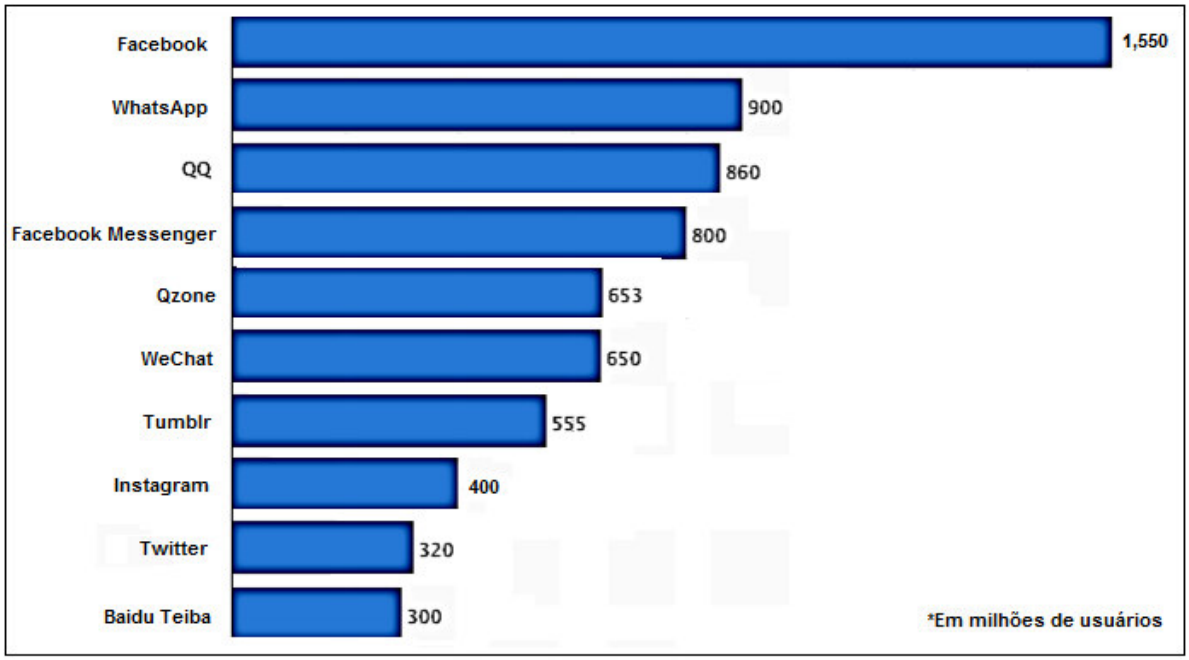


Figura 2.3: Ranking das redes sociais por usuários registrados no mundo
Fonte:Statista (2016)

Para Capra (2008), as redes sociais podem operar em diferentes níveis, como, por exemplo, redes de relacionamentos (*Facebook, Myspace*[6]*, Twitter*), redes profissionais (*LinkedIn*[7]), redes comunitárias (redes sociais em bairros ou cidades), redes políticas, permitindo analisar a forma como as organizações desenvolvem a sua atividade, como os indivíduos alcançam os seus objetivos ou medir o capital social – o valor que os indivíduos obtêm da rede social.

2.2.2 O Twitter

O *Twitter* é uma das redes sociais mais populares da atualidade, contando com mais de 320 milhões de usuários, responsáveis por trocarem e compartilharem mais de 500 milhões de *tweets*[8] diariamente Protalinski (2013).

Com características bem peculiares, o *Twitter* é um *microblog* [9] que permite que seus usuários leiam postagens pessoais de outros usuários. Uma das particularidades que merece ser observada é que, diferente de outras redes sociais, como o *Facebook*, exemplificativamente, permite montagem de redes por assimetria, ou seja, sem necessidade de autorização da outra parte para que possa ser seguido. A Figura 2.4 mostra um exemplo de perfil no *Twitter*.

[6]https://myspace.com
[7]https://www.linkedin.com
[8]Mensagens com no máximo 140 caracteres próprias do twitter
[9]Um tipo de blog que permite a escrita e atualização de textos curtos, geralmente inferior a 200 caracteres

Figura 2.4: Exemplo de um perfil no *Twitter*

As especificidades do *microblog*, no que tange a elementos hipertextuais ou não, exigem uma "alfabetização" do usuário e moldam a comunicação via *Twitter*, @(menção direta), #(hashtag) e *Retweet* ou RT. Estes elementos, somados a limitação de 140 caracteres de cada publicação, foram responsáveis pelo surgimento de uma nova "microsintaxe", característica dessa rede social Lemos e Santaella (2010).

Em uma pesquisa realizada por Java (2007), ficou evidente o grande interesse por parte dos usuários em compartilhar e consumir informações através do *Twitter*. Para os autores, o *microblog* atende a uma demanda atual da sociedade em rede, por uma comunicação mais rápida e objetiva, que tende a diminuir a necessidade de tempo e pensamento investido para a geração de conteúdo.

O *Twitter* possui uma API (*Application Programming Interface*) própria e aberta, que facilita o acesso à sua base de dados de mensagens. Através de argumentos de pesquisa, como georeferenciamento, datas e palavras-chave, dentre outras configurações, dados podem ser extraídos para várias finalidades. Uma delas, que vem sendo bastante utilizada recentemente é a de extração de conhecimento, como é o caso da mineração de opinião.

Neste trabalho será utilizado largamente o conceito de mineração de opinião, por se tratar de ponto de extrema relevância para o alcance dos fins desejados.

2.3 Descoberta do Conhecimento em Base de Dados

A internet é hoje um grande repositório de dados[10], que, muitas vezes, se analisados isoladamente, não representam informação[11], tampouco, conhecimento[12]. Nesse contexto, o processo de descoberta do conhecimento surge no intuito de extraí-lo de fontes de dados não estruturados através da aplicação de suas várias etapas.

Nesta seção serão abordados conceitos relacionados ao processo de Descoberta de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* - KDD) como parte da metodologia que será aplicada nesta pesquisa.

Para Fayyad et al. (1996), a descoberta de conhecimento em banco de dados (Knowledge Discovery in Databases - KDD) é um processo não trivial de identificação de padrões embutidos nos dados que sejam válidos, novos, potencialmente úteis e compreensíveis.

A descoberta de conhecimento se caracteriza por ser um processo complexo e que se destina a extrair informações em grandes volumes de dados, partindo da utilização de uma sequência lógica de etapas, como mostra a Figura 2.5, em que em cada uma delas se vê desempenhada uma função específica.

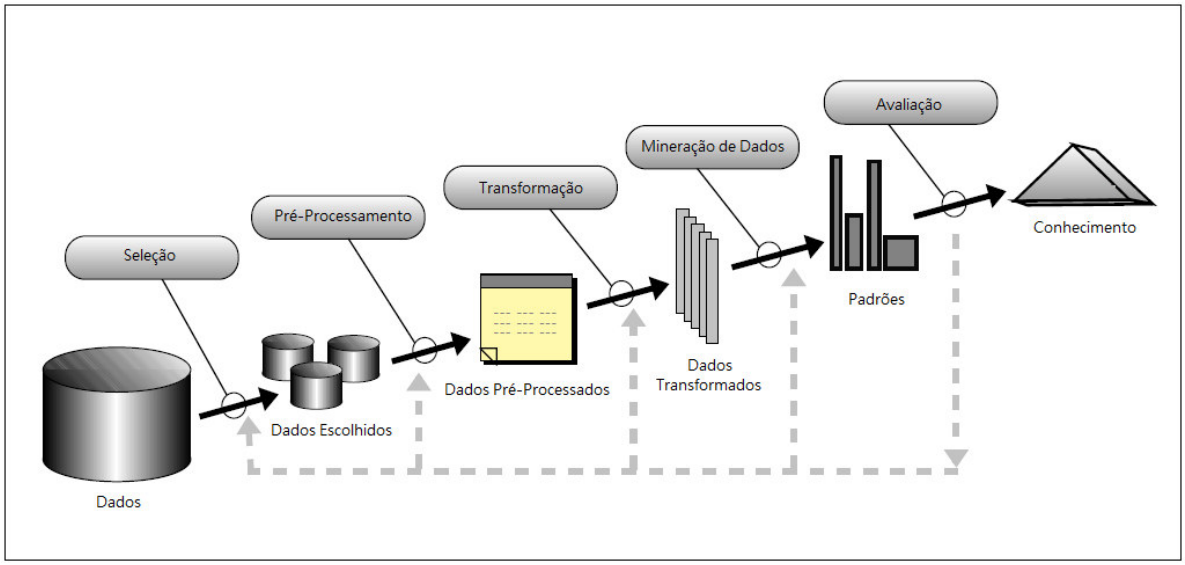


Figura 2.5: Processo KDD Fayyad et al. (1996)

[10] Elemento que representa eventos ocorridos na empresa ou circunstâncias físicas, antes que tenham sido organizados ou arranjados de maneira que as pessoas possam entender e usar Palmisano e Rosini (2003).

[11] Dado configurado de forma adequada ao entendimento e à utilização pelo ser humano Palmisano e Rosini (2003).

[12] É a capacidade, adquirida por alguém, de interpretar e operar sobre um conjunto de Informações Palmisano e Rosini (2003).

A Tabela 2.2 resume cada uma das etapas ilustradas na figura acima, correspondentes ao processo KDD.

Tabela 2.2: Etapas do processo KDD

ETAPA	**DESCRIÇÃO DO PROCESSO**
Seleção	Seleciona na base o conjunto de dados contendo todas as possíveis variáveis envolvidas (geralmente feita por um especialista)
Pré-Processamento/Limpeza	Remove os dados redundantes e inconsistentes
Transformação	Localiza características úteis que representem os dados de acordo com a tarefa escolhida, além de formata-los e armazena-los adequadamente para que os algoritmos de mineração de dados possam ser aplicados.
Mineração dos dados	Aplica os algoritmos para extração de conhecimentos dos dados
Avaliação/Interpretação	Apresenta as informações resultantes do processo de mineração
Conhecimento	Apresenta o conhecimento

Fonte: Adaptada do processo KDD Fayyad et al. (1996)

Para esta pesquisa será utilizado o processo KDD para extrair conhecimento sob a forma de sentimentos, utilizando como fonte de dados o *Twitter*, em que a grande quantidade de postagens sobre os mais diferentes assuntos, expressando diversos tipos de emoções, pode ser vista como uma fonte de informação em seu estado bruto, necessitando ser "lapidada" para que possa refletir o verdadeiro sentimento nela contido.

Na próxima seção serão mostrados de forma resumida, aspectos que envolvem a mineração de dados sendo uma das etapas mais importantes do processo KDD.

2.3.1 Mineração de Dados

A quantidade de dados gerados nas mais diferentes áreas do conhecimento tem crescido de maneira espantosa. Este crescimento exponencial gera não somente o desafio de armazenamento e gerenciamento do grande volume de dados (*Big Data*), mas também de como analisá-los e extrair conhecimento relevante Bakshi (2012);Demirkan e Delen (2012);Fan e Liu (2013).

Neste sentido, diversos modelos computacionais vêm sendo desenvolvidos com o intuito de simplificar o entendimento da relação entre as variáveis em grandes conjuntos de dados brutos. Algoritmos de Mineração de Dados podem auxiliar na descoberta de conhecimento, entretanto, estes algoritmos geralmente necessitam fazer uma leitura de toda a base de treinamento para obter as estatísticas necessárias para otimizar os parâmetros dos modelos, processo que requer

computação intensiva e acesso frequente aos dados em larga escala WU et al. (2014).

A mineração de dados tem como grande impulsionador os anseios da sociedade pela informação e, para isso, busca extrair conhecimentos fazendo uso de grandes volumes de dados, como os encontrados nas redes sociais, tendo como objetivo a extração de informações úteis que auxiliarão na tomada de decisão.

2.3.1.1 Mineração de Dados em Banco de Dados(*Data Mining*)

A mineração de dados, ou *Data Mining*, é parte integrante do processo KDD, sendo responsável pela seleção dos métodos utilizados na busca dos padrões ou similaridades entre os dados, de forma que sejam representativos e úteis para o domínio.

Mineração de Dados pode ser entendida como o processo de extração de informações, sem conhecimento prévio, de um grande banco de dados e seu uso para tomada de decisões. É uma metodologia aplicada em diversas áreas que usam o conhecimento, como empresas, indústrias e instituições de pesquisa Helena e Ângela C. (2003).

Extrair conhecimento de grandes bancos de dados é uma tarefa desafiadora, por isso a mineração de dados combina métodos e ferramentas de várias áreas, tais como aprendizagem de máquina, estatística, banco de dados, sistemas especialistas e visualização de dados, conforme ilustra a Figura 2.6, Cratochvil (1999).

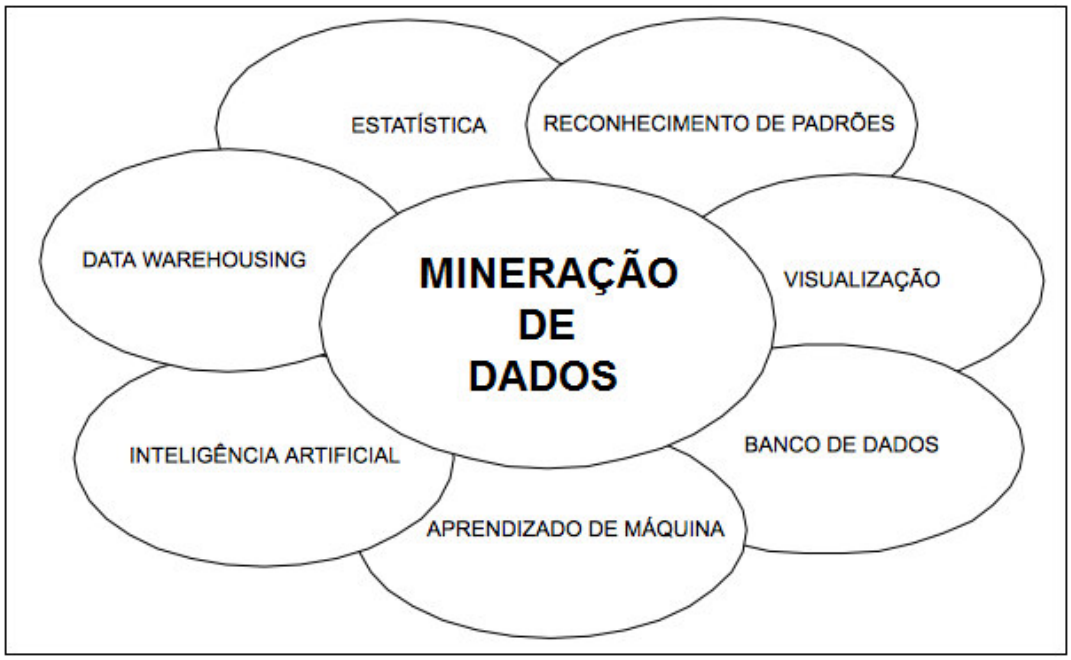


Figura 2.6: Áreas envolvidas na mineração de dados Neves (2003)

Segundo Cabena et al. (1997), a mineração de dados consiste no processo de extrair informação implícita, previamente desconhecida e potencialmente útil, desde as grandes bases de dados, usando-as para a tomada de decisão. Ainda consoante referendado autor, que analisa de uma perspectiva de banco de dados, descobrir regras, padrões globais e relacionamentos é um dos principais objetivos da mineração de dados.

Já Fayyad et al. (1996), sob uma ótica de aprendizado de máquina, define mineração de dados como sendo um passo no processo de descoberta de conhecimento, o qual consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta, que, sob certas limitações computacionais, produzem um conjunto de padrões de certos dados.

A mineração de dados pode ter suas técnicas aplicadas a tarefas[13] como classificação, estimativa, associação, agrupamento e sumarização. Na Tabela 2.3, pode-se verificar com mais detalhes cada tarefa.

Tabela 2.3: Tarefas realizadas por técnicas de mineração de dados(Baseado em:Dias (2001))

TAREFA	**DESCRIÇÃO**	**EXEMPLO**
Classificação	Constrói um modelo de algum tipo que possa ser aplicado a dados não classificados a fim de categorizá-los em classes	Identificar a melhor forma de tratamento de um paciente
Estimativa (Regressão)	Usada para definir um valor para alguma variável contínua desconhecida	Estimar o valor em tempo de vida de um cliente
Associação	Usada para determinar quais itens tendem a serem adquiridos juntos em uma mesma transação	Determinar quais os produtos costumam ser colocados juntos em um carrinho de supermercado
Agrupamento (Clustering)	Processo de partição de uma população heterogênea em vários subgrupos ou grupos mais homogêneos	Agrupar clientes por região do país
Sumarização	Envolve métodos para encontrar uma descrição compacta para um subconjunto de dados	Derivar regras de síntese

2.3.1.2 Mineração de Dados na Web(*Web Mining*)

A Web [14] é um universo heterogêneo que cresce de maneira exponencial e não estruturada. Por sua natureza dinâmica, milhares de páginas e conteúdos, de uma forma geral, surgem e desaparecem com a mesma facilidade a cada dia. Nesse contexto, ter acesso a informação de maneira clara e rápida não é tarefa simples. Nesse ambiente, a mineração de dados tem três enfoques para sua aplicação: Mineração de Conteúdo, Mineração de Estruturas da Web e Mineração de Uso da Web.

A Mineração na Web pode ser definida como a utilização de técnicas de mineração de

[13] Refere-se a qual tipo regularidades ou categoria de padrões deseja-se encontrar.
[14] Rede que conecta computadores por todo mundo, o mesmo que *World Wide Web* (WWW)

dados para a recuperação automática, extração e avaliação de informação para a descoberta de conhecimento em documentos e serviços da Web Pal et al. (2000). A Figura 2.7 ilustra esse processo.

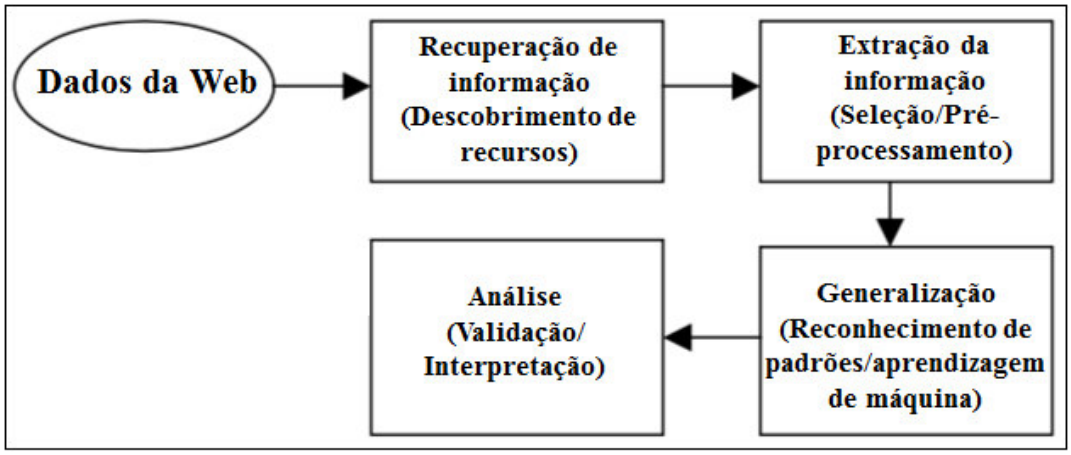


Figura 2.7: Sub-tarefas da mineração de dados na Web Pal et al. (2000)

Na Web, diferentemente da mineração tradicional, caracteriza-se pela existência de vínculos de hipertexto entre os seus documentos. Os ditos vínculos de hipertexto são uma rica fonte de informação a ser explorada, pois, dentre outras coisas, ajudam no processo de ranqueamento de páginas pelos motores de busca e na identificação de micro-comunidades na Web. Embora em ambiente diferente, agora na Web, convém ressaltar que a metodologia de descoberta do conhecimento segue os mesmos princípios descritos anteriormente.

2.3.1.3 Mineração de Dados Textuais (*Text Mining*)

Com o crescimento da Web e das redes de computadores de uma maneira geral, os documentos se tornaram um dos principais meios de armazenamento de informação. Uma grande quantidade de toda informação disponível atualmente encontra-se sob a forma de textos ou documentos não estruturados ou semiestruturados, tais como livros, manuais, revistas, artigos, e-mails e similares. Sendo assim, existe um grande gargalo na recuperação dessa informação que a mineração de texto tem procurado solucionar.

A mineração de texto (*Text Mining*), ou mais precisamente a Descoberta de Conhecimento em Textos (*KDT - Knowledge Discovery Text*), é um processo KDD, que consiste em utilizar técnicas de análise e extração de dados a partir de textos, frases ou palavras que se apresen-

tam normalmente de forma não estruturada. Este processo requer a aplicação de algoritmos computacionais em busca de padrões úteis e que normalmente não poderiam ser recuperados utilizando os métodos convencionais de recuperação de informação.

Para Gupta e Lehal (2009), a mineração de texto pode ser entendida com um processo de extração e obtenção de informações de qualidade diferenciada a partir de textos que se encontram em linguagem natural. O autor também comenta sobre a diferença da origem dos dados, no sentido de que, enquanto a mineração de dados obtém a informação de fontes estruturadas, a mineração de texto tem sua matéria-prima, originária de dados não estruturados ou semiestruturados.

Suas principais contribuições estão relacionadas com a busca de informações contidas em textos, em que as formas convencionais não obtêm grandes resultados, sendo necessária a utilização de técnicas mais específicas para a extração desse conhecimento armazenado sob o formato de textos.

A mineração em textos ganhou grande notoriedade com o crescimento da internet e das evoluções na área de linguística computacional (PLN). Seus benefícios e aplicações podem se estender a qualquer domínio que utilize textos Loh (2001).

A Figura 2.8, mostra o processo KDT e suas etapas.

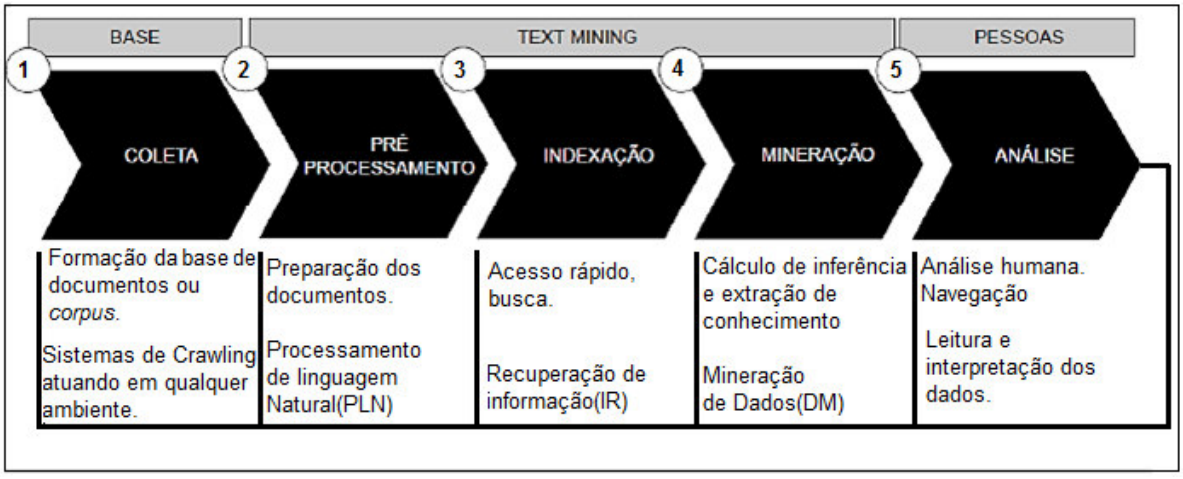


Figura 2.8: Etapas do processo de mineração de textos Aranha e Passos (2007)

A Tabala 2.4 apresenta as fases ilustradas acima e um resumo do que foi dito por Filho et al. (2010) sobre cada uma delas.

Tabela 2.4: Etapas do Processo KDT(Baseado em:Filho et al. (2010))

ETAPA	**DESCRIÇÃO DO PROCESSO**
Coleta	É formada a base textual para extração do conhecimento, também chamada de *corpus*
Pré-Processamento/Limpeza	Tem por objetivo formatar os textos coletados e deixá-los homogêneos
Indexação	Responsável pela organização dos termos resultantes das fases anteriores, visando facilitar a recuperação da informação
Mineração dos dados	Responsável pela inferência dos algoritmos e desenvolvimento dos cálculos para extração do conhecimento e descoberta dos padrões úteis
Análise dos resultados	Nesta etapa final, os dados resultantes do processo serão analisados por especialistas ou interessados no domínio, para que auxiliem no processo de tomada de decisão

O capítulo 4 deste trabalho abordará na prática essas etapas.

2.3.1.4 Processamento de Linguagem Natural (PLN)

Como dito na seção anterior sobre mineração de texto, existe uma grande quantidade de dados armazenados sob a forma de textos não estruturados, cujo paradigma tradicional da computação tende a criar grandes entraves na recuperação dessa informação. O processamento de linguagem natural ou linguística computacional surge nesse contexto como uma ferramenta capaz de obter informação partindo de textos e frases, muitas vezes dispersas e imprecisas, em linguagem natural.

Normalmente os computadores estão aptos a interpretar linguagens padronizadas, precisas e lógicas, como as voltadas a programação, no entanto ao se depararem com a linguagem natural, repleta de ambiguidades, gírias e sarcasmos, têm sua eficiência, no que se refere a recuperação de informação e conhecimento, reduzida ou imprecisa.

Quando se fala em PLN, seja do ponto de vista da compreensão humana ou computacional, percebe-se conceitos comuns, pois ambos os casos necessitam conhecer e manter informações fonéticas, morfológicas, sintáticas, semânticas, pragmáticas e um conjunto de palavras suportadas por tal linguagem. Abaixo estão descritos de forma mais detalhada esses níveis de compreensão da linguagem natural segundo Jurafsky e Martin (2000)

- **Fonético** - relacionamento das palavras com os sons da língua;

- **Morfológico** - construção das palavras a partir de unidades de significado primitivas e de como classificá-las em categorias morfológicas (conjugação, declinação etc.);
- **Sintático** - relacionamento das palavras entre si, seu papel estrutural nas frases; trata da disposição das palavras numa sentença;
- **Semântico** - estabelece correspondência entre situações do mundo real ou do mundo possível e as estruturas reconhecidas ao nível sintático;
- **Pragmático** - interpreta essas situações no contexto mais geral de uma troca de informações;
- **Discurso** - o estudo de unidades linguísticas maiores do que um enunciado [15].

Ressalta-se que boa parte da dificuldade encontrada pela linguística computacional está na incapacidade de compreender todos os níveis com a mesma clareza, sobretudo a sintaxe, a semântica e a pragmática, que requerem grandes esforços computacionais.

O PLN tem diversas aplicações práticas, dentre elas destacam-se a seguir as citadas, com maior relevância baseadas nos estudos de Gonzalez e Lima (2003) e Dias da Silva (2007):

- Recuperação de informação (busca e filtragem);
- Classificação de texto (muito usado na mineração de opinião);
- Estruturação de hipertextos (geração e manutenção de páginas de links);
- Tradução (automática e semi-automática);
- Auxiliar na auditoria de documentos (verificação orto-gramatical);
- Sistemas de diálogos (ajuda on-line, tutores inteligentes);
- Mineração de opinião (extração do sentimento).

Dentre estas aplicações, a mineração de opinião é uma das mais recentes e promissoras, tendo como meta o processamento de textos com opinião, sentimento e subjetividade Pang e Lee (2008). Analisar estes aspectos em fontes textuais é um dos objetivos deste trabalho.

O ramo do PLN ainda é cercado de muitas expectativas e decepções, mas indiscutíveis avanços que certamente representarão o início de um grande progresso no campo da inteligência artificial.

[15] É uma unidade de fala. Pode se apresentar como: palavra, frase ou oração. Caracteriza-se pela entonação e que enfatiza o significado que se pretende transmitir

2.3.1.5 Aprendizado de Máquina e SVM

Nesta seção, para o melhor entendimento da pesquisa, será feita uma breve conceituação sobre aprendizado de máquina e SVM (*Support Vector Machine*).

O aprendizado de máquina (*machine learning*) é uma das técnicas utilizadas na mineração de dados que utiliza conceitos de inteligência artificial para desenvolver modelos que possam "aprender" através de exemplos.

Seu processo fundamenta-se em um princípio da inferência chamado indução, que consiste em obter conclusões genéricas, tomando como base um conjunto de exemplos. De uma forma geral, um conceito é aprendido pela indução, tendo como pilar os exemplos utilizados, no entanto seu resultado pode ou não representar a verdade Rezende et al. (2003). O aprendizado indutivo ocorre a partir de um raciocínio que tem como base exemplos fornecidos por um processo externo ao sistema de aprendizado.

O principal objetivo desta técnica é encontrar padrões e regras gerais em grandes quantidades de dados, de modo que permita a extração de informações úteis. As técnicas de aprendizado de máquina se dividem em dois tipos: aprendizado supervisionado e aprendizado não supervisionado Tan et al. (2006). A Figura 2.9 ilustra essa divisão e suas ramificações.

Figura 2.9: Hierarquia do Aprendizado Indutivo Rezende et al. (2003).

Este trabalho fará uso da aprendizagem supervisionada, também definida a seguir:

- **Aprendizado Supervisionado** - Neste tipo de aprendizado, um banco de dados contendo um modelo de classificação é construído, baseado em um conjunto de instâncias de classes para treinamento. Tomando como base este modelo predefinido, o processo deverá ser capaz de prever associações de classe para novas instâncias, a partir de suas características. Neste método, a figura do especialista é indispensável, pois ele é o responsável pela classificação das instâncias utilizadas para treinamento Williams et al. (2006).

- **Aprendizado não Supervisionado** - Este aprendizado caracteriza-se pela incerteza sobre o resultado (saída), pois nele não se conhece, a princípio, a classificação das instâncias que é feita por critérios adotados pelo próprio algoritmo. Neste método, é dispensada a figura do especialista o qual tem sua função assumida pelo algoritmo do próprio método. Este tipo de aprendizado é bastante utilizado para resolver problemas ligados a agrupamentos.

Fundamentada na Teoria da Aprendizagem Estatística, a SVM foi concebida por Vapnik (1995), com a finalidade de resolver problemas de classificação, tendo sido usufruída com sucesso em aplicações de reconhecimento de padrões em diversas áreas.

De uma maneira simplificada, o funcionamento de uma SVM pode ser descrito como segue: dadas duas classes compostas por pontos, a SVM tem como finalidade determinar o hiperplano que separe os pontos da melhor forma possível, colocando o maior número de pontos comuns a uma classe do mesmo lado, e ao mesmo tempo que maximiza a distância de cada classe em relação ao hiperplano. A distância que compreende a classe a um hiperplano é a menor distância entre ele e os pontos dessa classe, denominado de margem de separação. Esse hiperplano gerado pela SVM é formado por um subconjunto dos pontos que compõem as classes envolvidas, que é chamado “vetores suporte”. As Figuras 2.10 e 2.11 ilustram o processo descrito.

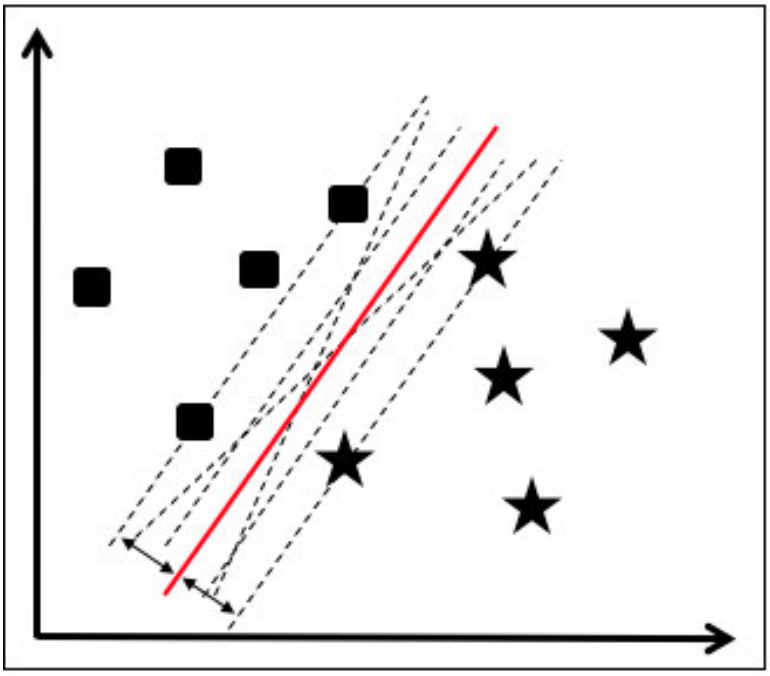


Figura 2.10: Hiperplano ótimo (Baseado em:(Cristianini e Shawe-Taylor, 2000))

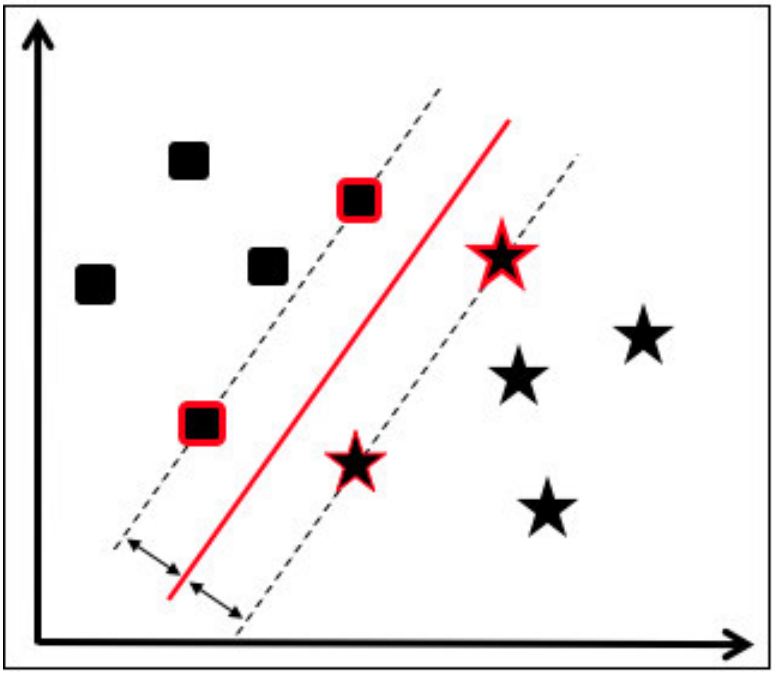


Figura 2.11: Vetor de Suporte (Baseado em:(Cristianini e Shawe-Taylor, 2000))

Neste trabalho será utilizado o aprendizado supervisionado, onde dado um conjunto de exemplos rotulados sob a forma $(x_i; y_i)$, em que x_i representa um exemplo e y_i o seu rótulo, deve-se produzir um modelo ou classificador que seja capaz de predizer precisamente o rótulo de novos dados de entrada. Esse processo de indução de um modelo a partir de uma amostra

de dados é conhecida como treinamento. O classificador obtido também pode ser visto como uma função f, a qual recebe um dado x e fornece uma predição y.

2.3.2 Mineração de Opinião

Podemos definir a opinião em um texto opinativo como sendo a resultante resumidamente o caráter

A exemplo de outros segmentos ligados a área de mineração de dados, a mineração de opinião também teve grande impulso com o crescimento dos dados disponíveis na Web, sobre tudo nos últimos anos. A Web tornou-se um local potencialmente rico em informações, por ser uma espécie de "concentrador" natural do que acontece no mundo real. Empresas, clientes, usuários de redes sociais, blogs, microblogs, entre outros, contribuem de forma quase que frenética com seus comentários, opiniões, críticas e sugestões sobre os mais diversos assuntos.

Com o intuito de classificar as opiniões (positiva, negativa ou neutra) contidas nos textos dos usuários sobre determinada entidade e transformá-las em informação, surge uma área promissora, a mineração de opinião, também conhecida como análise de sentimentos, que faz uso do processamento de linguagem natural, em busca do sentimento [16] contido em textos opinativos [17] e que nem sempre são de fácil percepção.

Paralelamente aos desafios trazidos pela grande quantidade de dados produzidos na contemporaneidade, pode-se dizer que a mineração de opinião é um segmento em plena expansão, vindo boa parte desse crescimento, de uma espécie de relação mútua entre o grande volume de dados gerados e a oportunidade de poder extrair conhecimento de algo que isoladamente não representa informação.

Segundo Pang e Lee (2008), a mineração de opinião é um ramo da mineração de textos, que se preocupa em classifica-los não por tópicos, e sim pelos sentimentos ou opinião contida em determinado documento. Geralmente associado à classificação binária entre sentimentos positivos e negativos, o termo é usado de uma forma mais abrangente para significar o tratamento computacional de opinião, sentimento e subjetividade em textos.

Já para Liu (2010), a mineração de opinião, ou *Opinion Mining*, é o estudo computacional de opiniões, sentimentos e emoções expressadas em texto. A informação textual pode ser classificada em dois tipos principais: fatos e opiniões. Os fatos são expressões objetivas sobre entidades, eventos e as suas propriedades. As opiniões são geralmente expressões que descrevem

[16] Doravante também podendo ser chamado de humor

[17] Textos que qualificam os aspectos. Na maioria das vezes são os adjetivos e advérbios

os sentimentos e avaliações das pessoas em relação a determinadas entidades, eventos e suas respectivas propriedades.

Frisa-se que a mineração de opinião tem sido uma das áreas de pesquisa mais ativas no campo do PLN e tem como principal desígno obter e formalizar a opinião e o conhecimento subjetivo em documentos não estruturados (textos), para posterior análise dentro de um domínio específico Liu (2012).

A opinião sempre teve um papel de destaque na sociedade, em quase todas as suas atividades, pois influencia diretamente nas escolhas e comportamentos. A consulta de uma opinião normalmente antecede uma tomada de decisão, podendo ser feita ou não por um especialista Liu (2012). Ele também define formalmente a opinião como sendo uma quíntupla $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, em que:

- e_i - é nome de uma entidade;
- a_{ij} - é o aspecto da entidade e_i. Um aspecto também é denominado tópico;
- s_{ijkl} - é a opinião sobre a_{ij} da entidade e_i;
- h_k - é a entidade que expressou a opinião, também chamado de fonte de opinião ;
- t_l - é o tempo no qual foi expressa por h_k.

A opinião expressa por s_{ijkl}, sobre uma entidade ou aspecto é medida em termos de uma polaridade, que pode ser classificada em: positiva, negativa ou neutra Tsytsarau e Palmanas (2012)

Uma vez classificada a polaridade do sentimento, seu resultado pode ser de grande valia para distintas áreas. Na Tabela 2.5 seguem ramos e uma breve descrição de sua utilização.

Tabela 2.5: Aplicações para mineração de opinião(Baseado em:Liu (2012))

ÁREA	**APLICAÇÃO**
Política	Para medir a popularidade de um candidato a determinado cargo público;
Indústria	Para avaliar a aceitação por parte dos consumidores a um determinado produto;
Bolsa de valores	Para medir o sentimento coletivo sobre determinado ativo em negociação na bolsa.

Embora suas aplicações sejam extremamente úteis na tomada de decisão, extrair sentimento em fontes textuais é ainda uma tarefa complexa, pois o PLN, que é parte integrante do processo de mineração de opinião, muitas vezes se depara com expressões cercadas de neologismos, ironia e outras variações linguísticas que dificultam a correta extração do sentimento. Para esta pesquisa, especificamente, será utilizada a mineração de opinião para a área financeira, mais especificamente no domínio bolsa de valores.

2.3.2.1 Níveis de Mineração de Opinião

O objetivo da mineração de opinião é identificar as sentenças que contém as opiniões a serem comparadas, fazendo uso, exemplificativamente de advérbios como "pior que", "melhor que" etc., e, assim extrair a entidade referida daquela opinião específica Feldman (2013). Isso, segundo Liu (2012), pode ser aplicado em três níveis: documento, sentença ou por entidade e aspecto. Abaixo uma síntese sobre cada nível:

- **Nível de Documento** - A tarefa é classificar a opinião geral do documento como um sentimento positivo ou negativo Turney (2002). Esse tipo de análise suporta avaliação que expressem opiniões sobre apenas uma entidade, não sendo possível, portanto analisar textos que fazem a comparação de várias entidades;

- **Nível de Sentença** - Tem como tarefa determinar a cada sentença se o sentimento expresso é positivo, negativo ou neutro, valendo registrar que neutro significa sem opinião. Este nível de análise está intimamente relacionado à classificação de subjetividade Wiebe (2002), que diferencia sentenças objetivas, que expressam informações concretas, de sentenças subjetivas, as quais expressam visões pessoais;

- **Nível de entidade e aspecto** - Este é o mais complexo dos níveis; nele uma sentença pode ser julgada por várias entidades e também pode conter múltiplos sentimentos associados a ela. Sendo assim, pode-se descobrir o que exatamente pessoas aprovam ou desaprovam, diferentemente das análises em níveis de documentos e de sentenças, que limitam a análise nesse aspecto. Esse modelo de avaliar a opinião para cada entidade é muito utilizado para fóruns de discussões e *reviews*.

Os métodos mais utilizados para detecção de sentimentos em sentenças na atualidade podem ser divididos em duas classes: os métodos léxicos e os baseados em aprendizado de máquina. Os primeiros utilizam listas e dicionários de palavras associadas a sentimentos específicos. Os métodos baseados em aprendizado de máquina geralmente dependem de bases de dados rotuladas para treinar classificadores Pang e Vaithyanathan (2002), o que pode ser considerado uma desvantagem devido ao alto custo na obtenção de dados rotulados.

Nesta pesquisa será adotado o nível de sentença, por ser o mais indicado para utilização em redes sociais e textos curtos, como os do *Twitter*, que tem a limitação de 140 caracteres.

2.4 Síntese

Neste capítulo, foi feita a fundamentação teórica sobre os conceitos relevantes para a pesquisa, tais como: redes sociais, mercado financeiro e o processo de descoberta de conhecimento em base de dados (KDD). Este ultimo pode ser considerado como um roteiro que norteará todo o trabalho.

No capítulo seguinte serão apresentados trabalhos que contribuíram para o enriquecimento deste estudo, fornecendo informações sobre suas metodologias e temáticas abordadas.

Capítulo 3

Trabalhos Relacionados

Uma vez definida a problemática envolvida na presente pesquisa, iniciar-se-á uma investigação em busca de trabalhos relacionados com a temática em estudo. Neste capítulo serão elencados alguns trabalhos, cujos temas estejam na mesma linha desta pesquisa. Serão feitos alguns comentários e comparativos entre as técnicas utilizadas.

3.1 Visão Geral

A predição do mercado financeiro é um assunto recorrente, sendo que várias são as técnicas utilizadas nessa tentativa, algumas com maior eficiência e outras que não repetem o mesmo sucesso. No entanto, trabalhos relacionados à utilização de dados em fontes textuais e mineração de opinião para esta finalidade não são muito comuns e fazem parte de uma nova vertente no segmento de predição da bolsa, fazendo uso de técnicas de inteligência artificial. A maioria dos trabalhos estudados utiliza metodologias semelhantes em alguns pontos.

Etapas que fazem parte do processo KDD são imprescindíveis por se tratarem de um ciclo sequencialmente lógico na descoberta do conhecimento. Convém ressaltar que a utilização da mineração de opinião como ferramenta estratégica no processo de tomada de decisão pode ter várias aplicações, como já mencionado no capítulo anterior, sendo que esta pesquisa se limitará a avaliar trabalhos relacionados à mineração de opinião em redes sociais e, mais especificamente, com à finalidade de predição sob o ponto de vista financeiro.

A mineração de opinião utilizando as redes sociais, mais especificamente o *Twitter*, que nesse contexto faz o papel de uma espécie de termômetro social, tem ganhado grande notoriedade. A maioria desses trabalhos se apoia nos estudos de Pang e Lee (2008) sobre monitoramento

em redes sociais e mineração de opinião para descoberta de conhecimento e tem demonstrado sua eficácia quando utilizada para fins de predição. A seguir estão relacionados alguns desses trabalhos, seus objetivos, metodologias e resultados.

3.2 Previsão de Arrecadação com Bilheteria de Filmes

O lançamento de um filme é certamente um evento cercado de sentimentos e emoções, como, por exemplo, a ansiedade, notadamente sob dois aspectos, a dizer, o do entretenimento, nos amantes da chamada sétima arte e o aspecto econômico, representado por aqueles que buscam retornos financeiros com esse mercado. No intento de relacionar as duas vertentes, o trabalho descrito a seguir obteve bons resultados.

Em seu trabalho intitulado "*Predicting the Future with Social Media*", Asur e Huberman (2010) propuseram um modelo utilizando regressão linear, que, baseado nos comentários disponíveis no *Twitter* sobre determinado filme durante sua pré-estreia, busca estabelecer uma relação onde o filme bem comentado na rede terá seu sucesso refletido também nas bilheterias. Com isso, pretende predizer alguns indicadores importantes para o domínio: a arrecadação da primeira semana de bilheteria, o rendimento de todos os filmes lançados em um determinado período e a relação com o índice *Hollywod Stock Exchange*[1].

Para a vertente pesquisa, o conjunto de dados foi obtido utilizando *Twitter Search API*[2], coletados a cada hora. Como argumento de busca usadas palavras-chave presentes no título do filme, resultando em 2,89 milhões de *tweets* de 1,2 milhão de usuários, referentes a 24 filmes diferentes ao longo de 3 meses, período compreendido entre 20/11/2009 à 26/02/2010. Os autores utilizaram dois aspectos para predição: quantitativo, em que é levada em conta a quantidade de citações relacionadas ao filme e o baseado no humor expresso pelos usuários do *Twitter*.

Para a classificação do humor dos *tweets*, torna-se necessário o uso de um classificador, que neste caso específico foi o *DynamicLMClassifier* [3], o qual resulta em três sentimentos: positivo, negativo ou neutro.

Ao fim do trabalho, Asur e Huberman (2010) ressaltam o sucesso de suas predições, onde

[1] HSX - Hollywood Stock Exchange (Bolsa de Valores de Hollywood), é o mercado de entretenimento, onde você pode comprar e vender filmes, ações virtuais de celebridades, tudo através de uma moeda também virtual, diponível em http://www.hsx.com.

[2] API que permite consultas em relação aos *tweets* recentes ou populares a partir de um critério especificado. (https://dev.twitter.com/rest/public/search)

[3] http://www.alias-i.com/lingpipe

seu método foi mais eficaz do que o *Hollywood Stock Exchange*, que é uma referência no domínio. Observou-se forte relação entre a atenção a um determinado tópico dada pelos usuários e o desempenho da bilheteria no futuro, bem como a importância da apuração do sentimento dos *tweets* na melhora dos resultados. A Figura 3.1 ilustra os bons resultados da pesquisa, comparando a predição HSX x *Twitter*, constatando-se apenas pequenas variações entre as duas análises.

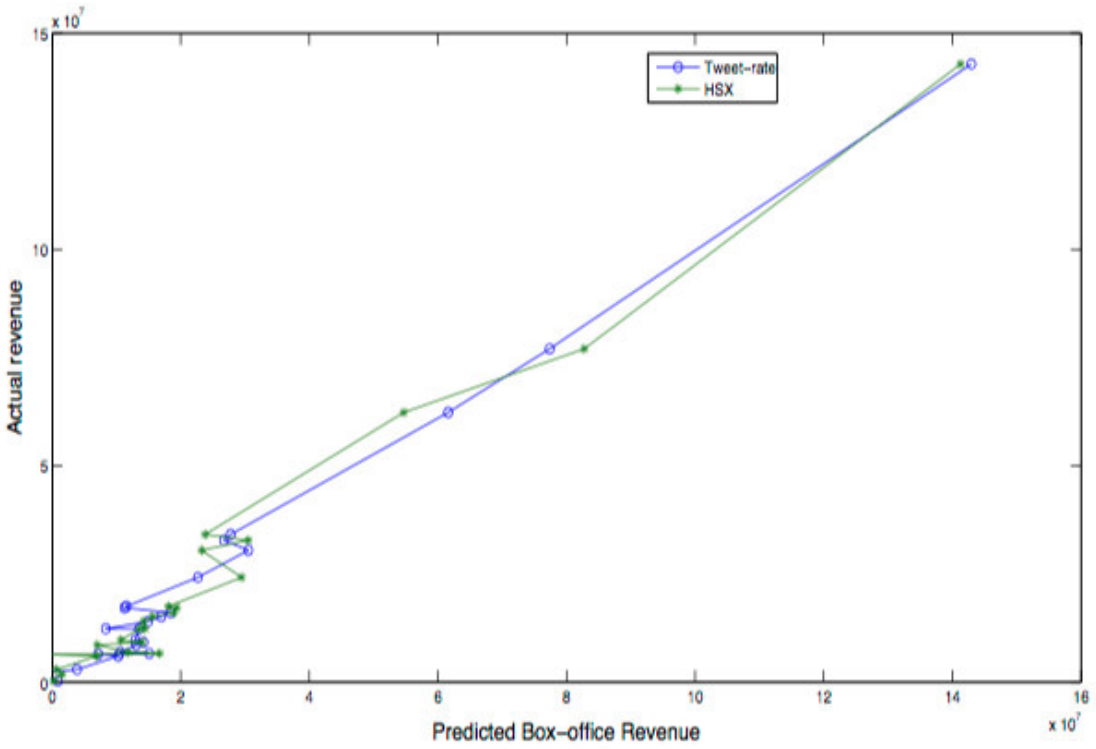


Figura 3.1: Predição X pontuação de bilheteria real usando *Twitter* e HSX Asur e Huberman (2010)

Segundo o autor, o estudo comprova que os resultados obtidos conseguiram superar os do HSX, havendo uma forte correlação entre a quantidade de atenção dada a um filme nas redes sociais e sua futura classificação.

3.3 Mineração de Opinião Voltada para Bolsa de Valores

No segmento predição da bolsa de valores, um dos trabalhos recentes que merece destaque é o de Bernardo (2014), com o título "*A Era de um Mercado Social: A Relação Entre o Twitter e o Mercado Acionista*". Sua pesquisa busca compreender a relação do *Twitter* com o mercado acionista, estabelece um vínculo entre dois aspectos importantes na sociedade, quais sejam, o

social e o financeiro, tendo como objetivo a predição dos ativos de um conjunto de empresas.

Para isso utiliza-se de duas técnicas para análise do sentimento nas postagens do *twitter*, uma léxica[4] e outra utilizando SVM, ambas almejando encontrar a maior acurácia nos resultados. Os sentimentos resultantes dessas duas análises são: positivo, negativo ou neutro.

Em sua metodologia, o autor referendado coletou dados através da *Twitter Search API*, entre os dias 11/02/2014 e 28/02/2014, em horário local compreendido de 13h30min às 21h00min, onde o mercado financeiro estava em plena atividade, resultando em 2.010.407 *tweets* coletados. Os dados foram obtidos em tempo real (*real-time*) e armazenados em banco de dados, tendo sido selecionadas para o estudo 16 empresas, dentre as quais: *Microsoft*, *Nike*, *LinkedIn*, *Amazon*, *Sony*.

Uma das características da pesquisa se limita a coletar as postagens somente na língua inglesa, por razões estratégicas, considerando que as empresas escolhidas para o estudo têm grande participação onde esse é o idioma local, e pelo fato dos léxicos também estarem nessa língua. Sendo assim, a etapa de tradução, em que pode haver ruídos e perda de informação, torna-se desnecessária.

Em outro processo, em que é evidente o KDD, em sua fase de limpeza dos dados, os mesmos tiveram suas inconsistências e informações desnecessárias retiradas, de modo que caracteres que porventura vinham a interferir no resultado foram suprimidos. Após essa etapa, 23% dos *tweets* foram descartados da amostra, por não atenderem aos requisitos.

Em seu primeiro teste utilizando a análise léxica, o autor recorreu à personalização de um dicionário já existente, utilizado no trabalho de Hu e Liu (2004), incluindo palavras e ícones que julgou relevantes, isso mediante alguns testes de ocorrência dessas palavras em seu *corpus* e uma adaptação que julgou natural, por originalmente ter sido utilizado para outro domínio. Utilizando este artifício, percebeu-se um ganho na precisão dos *tweets* analisados, que saiu de 0,67 para 0,68, embora pequena, mas representativa. O segundo teste, desta vez utilizando técnicas de aprendizado de máquina, consiste em: dado o *corpus*, o algoritmo irá aprender, tendo como base as palavras e expressões que constam nessa amostra. E uma vez treinado, servirá para classificar novas entradas. A precisão obtida nessa classificação foi de 81,71%, que representa um número bastante expressivo, cabendo ressaltar que o autor adotou uma metodologia, neste caso, inspirado no trabalho de Pang e Vaithyanathan (2002).

[4]É uma abordagem que utiliza-se de léxicos (dicionários) de sentimentos, que são conjuntos de palavras ou expressões de sentimento associadas a seu respectivo humor, previamente definido.

Em sua investigação, Bernardo (2014) também testa a relação Twitter/Mercado no que ele mesmo chama de "sentido inverso", ou seja, a capacidade que um segmento tem de influenciar o outro. Seus testes foram segmentados e agrupados nos seguintes intervalos: a cada 3 minutos, a cada 1 hora e a cada dia e obteve sentimentos diferentes para cada forma de agrupamento, assim como a seu poder preditivo sobre determinadas empresas também não foi o mesmo.

Por fim, o autor conclui em sua pesquisa que a capacidade de predição do *Twitter*, quando associada ao mercado financeiro, pode estar diretamente relacionada com a forma que os dados *tweets* são agrupados, bem como a característica da empresa as torna mais sensível a esse tipo de predição, sobretudo as empresas tecnológicas de sua amostra (LinkedIn, Cisco, Microsoft) e uma fraca relação no agrupamento diário utilizando essa metodologia, o que não acontece quando o intervalo testado é a cada 3 minutos.

3.4 Predição de Indicadores de Mercado Através do *Twitter*

Outro trabalho, ainda no segmento bolsa de valores, intitula-se "*Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"*" é o de Zhang et al. (2010), que, motivado pelo poder de predição implícito nos *tweets*, tenta prever indicadores do mercado de ações para bolsas de valores, como: Nasdaq, S&P 500 e Dow Jones.

O autor, baseado em sua fundamentação teórica, afirma que o estado emocional pode influenciar nas decisões, inclusive no mercado financeiro. Constata igualmente que a maioria dos *tweets* tem significado simples, em que é fácil capturar o tema principal. Sendo assim, opta por utilizar palavras como "medo","preocupação", "esperança", dentre outras, como identificadoras de emoção em um *tweet* e direciona sua coleta de informações para as postagens que contenham as referidas palavras emocionais.

A Tabela 3.1 demonstra o resultado da classificação, cujas palavras foram divididas em dois grupos: Positivo (Esperança, Feliz) e Negativo (Medo, Preocupação,Nervoso, Ansioso, Chateado, Positivo, Negativo). Esses dados foram coletados por um período de 6 meses, compreendido entre 30/03/2009 a 07/09/2009.

Tabela 3.1: Número de *Tweets* postados no período 30/03/2009 a 07/09/2009, adaptado de Zhang et al. (2010)

Emoção	**Média/Dia**	**Min/Dia**	**Máx/Dia**
Tweet	29758	8100	43040
Esperança	307	54	467
Feliz	260	37	1806
Medo	28	4	49
Preocupação	27	5	51
Nervoso	13	0	36
Ansioso	4	0	9
Chateado	14	2	25
Positivo	570	91	2204
Negativo	86	11	125

O trabalho se preocupou, em avaliar o tempo de reação do mercado ao que está sendo comentado no *Twitter*, mencionando a perda de informação nos dias não influenciáveis, que são os dias em que o mercado não está em operação e cujas postagens são irrelevantes para o estudo. Nesse cerne, quando parte para analisar *Dow Jones* de segunda-feira, por exemplo, os dados relevantes são os de domingo, ignorando-se sexta-feira e sábado.

Por fim, o autor conclui que quando as emoções no twitter estão em alta, isto é, quando os usuários expressam muita esperança e medo, normalmente no dia seguinte o *Dow Jones* sinaliza com uma queda. Já quando há menos esperanças, medo e preocupação nas postagens, o *Dow Jones* sobe. Ressalta também que as explosões emocionais monitoradas no *Twitter*, podem constituir um preditor do comportamento para o dia seguinte.

Zhang et al. (2010) deixa claro em sua pesquisa que seus resultados são preliminares e que podem ser aprimorados e complementados em pesquisas futuras.

3.5 Utilizando o Humor do *Twitter* para Predição de Ativos

Outro trabalho de destaque e com grandes contribuições para a área de mineração de opinião e predição no *Twitter* foi o de Bollen et al. (2011). Seu artigo "*Twitter mood predicts the stock market*", que também busca predizer o comportamento das bolsas de valores, realizando experimentos em sua pesquisa e procurando evidências que comprovem que o humor expresso no *twitter* influencia no comportamento da *Dow Jones Industrial Average* (DJIA) ao longo do

tempo.

O trabalho consiste na coleta de dados [5] do *Twitter* diariamente, no período entre de 28/02/2008 a 19/12/2008, resultando em 9.853.498 de *tweets* para algo em torno de 2.7 milhões de usuários.

Foram utilizadas para extração do humor duas ferramentas: *OpinionFinder* que resulta em dois estados do humor (positivo e negativo), e *Google-Profile of Mood States* (GPOMS), que tem como resultado 6 variações do humor (calma, alerta, Claro, Vital , gentil e feliz).

Após a fase de coleta os *tweets* foram armazenados e submetidos a um processo de pré-processamento ou limpeza, em que houve a retiradas de informações desnecessárias para a pesquisa, depois foram agrupados por data e considerados somente os que expressavam claramente o humor, com frases como: “eu sinto”, “Eu estou sentindo”, “eu não sinto”, “eu sou”.

As dimensões do humor GPOMS foram totalizadas por dias e seus resultados comparados com datas conhecidas no mesmo período, como ação de graças, eleições etc., concluindo que o humor resultante é típico destas datas.

Por derradeiro, visando a predição, o autor recorre a uma técnica baseada em rede neural *fuzzy*, denominada *Self Organizing Fuzzy Neural Network* (SOFNN), tendo em vista correlacionar as séries temporais resultantes da análise do *OpinionFinder* e GPOMS com o DJIA. A Figura 3.2 ilustra o diagrama com as três fases utilizadas na metodologia:

- criação e validação de OpinionFinder e GPOMS;
- uso da análise de causalidade de Granger para determinar correlação entre DJIA, OpinionFinder e GPOMS;
- formação de uma Rede Neural distorcida de auto-organização para prever valores DJIA com base em várias combinações de valores DJIA anteriores.

[5] É a fase da pesquisa em que se reúnem dados através de técnicas específicas.

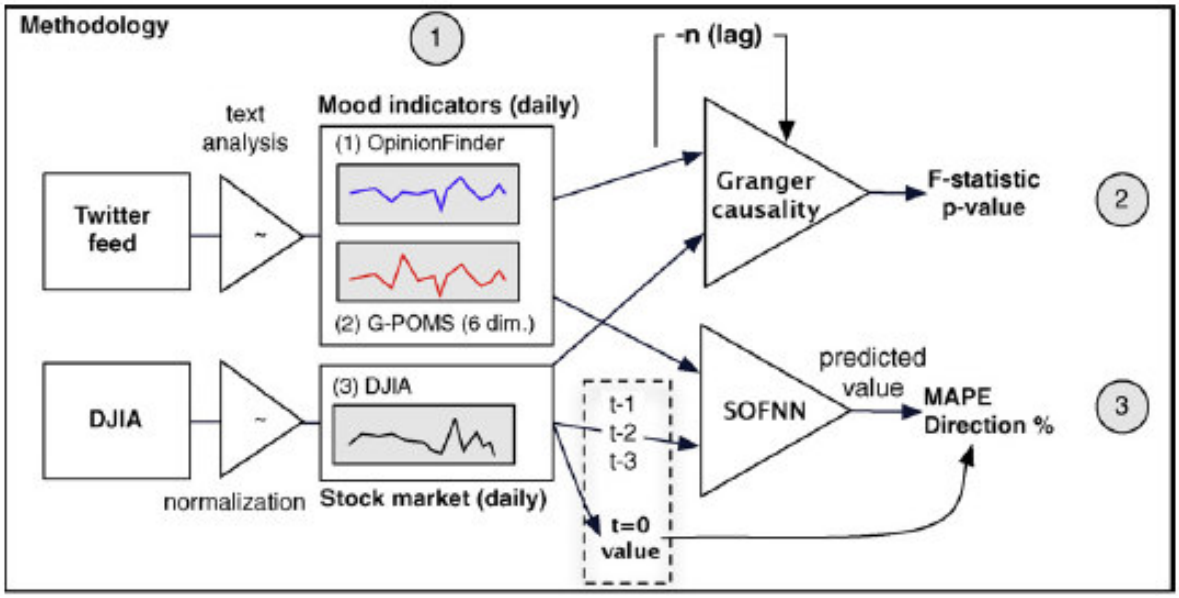


Figura 3.2: Diagrama delineando 3 Fases da metodologia e conjuntos de dados correspondentes Bollen et al. (2011)

Ao final de seus experimentos, o autor conclui que a abordagem que utiliza as opiniões (positiva e negativa) não obteve bons resultados, todavia, a que utilizou as emoções (calma e felicidade) demonstrou estar relacionada com o DIJA em alguns intervalos do período analisado. Outro aspecto conclusivo da pesquisa, segundo o autor, é que a polaridade da opinião é muito abrangente e oculta às emoções que são mais subjetivas.

3.6 Discussão

A predição de caráter financeiro independente da forma como se apresenta, seja na bolsa de valores, seja associada às vendas de um produto/serviço; ainda gera poucos trabalhos quando comparada a outras áreas afins. Dentre os trabalhos mencionados neste capítulo, serão demonstrados nessa seção alguns breves comentários sobre suas características.

Observou-se uma forte presença dos legados do trabalho de Pang e Lee (2008) sobre a mineração de opinião, em que sua pesquisa demonstra ser possível obter conhecimento a partir da análise e classificação de textos.

Outro ponto comum entre os trabalhos aqui comentados, é a utilização do *Twitter* como principal fonte de dados, ratificando seu potencial em refletir a opinião dos usuários através suas postagens. Etapas do processo KDD também são evidentes em todos os estudos, tanto na seleção, como no pré-processamento, transformação, mineração dos dados, interpretação,

até chegar ao conhecimento propriamente dito.

Por outro viés, alguns dos trabalhos tratam de maneira diferente a apuração do sentimento em sua metodologia. Nos trabalhos de Asur e Huberman (2010) e Bernardo (2014), o sentimento apurado varia entre positivo, negativo e neutro. Já para Zhang et al. (2010) este mesmo sentimento sofre uma espécie de agrupamento, onde há o positivo (Esperança, Feliz) e o negativo (medo, preocupação, nervoso, ansioso, chateado, positivo, negativo). Em Bollen et al. (2011), a apuração do humor tem duas abordagens, ora como positivo e negativo, ora apresentando 6 variações (calma, alerta, claro, vital , gentil e feliz). As variações na definição do humor encontradas em alguns dos trabalhos referenciados são justificados pela metodologia adotada em busca de uma nova abordagem ou melhoria nos resultados.

Quando da comparação dos estudos relacionados e o trabalho ora proposto, encontram-se alguns pontos comuns, principalmente nas etapas que envolvem o KDD, mencionadas anteriormente. Porém, embora de domínio semelhante, a metodologia adotada se diferencia dos demais trabalhos, algo que será demonstrado com maior riqueza de detalhes no próximo capítulo.

3.7 Síntese

Neste capítulo foram apresentados alguns trabalhos relacionados com a pesquisa em voga, trabalhos que na etapa de estado da arte foram de grande importância, fornecendo informações sobre o atual estágio das publicações científicas relacionadas ao tema escolhido.

Todos os estudos relacionados foram conclusivos e de grande valia, dando a sua contribuição científica dentro do que se propuseram, alguns com bons resultados, outros ainda necessitando aprimorar sua metodologia ou abordagem, mas um aspecto positivo comum que todos destacam é o grande potencial preditivo contido nas redes sociais, principalmente no *Twitter*, por apresentar características bem peculiares.

O próximo capítulo apresentará o trabalho proposto, em que será demonstrada a metodologia aplicada nesta pesquisa, fazendo uso do referencial teórico estudado até o momento.

Capítulo 4

Um Modelo de Predição de Bolsa de Valores Baseado em Mineração de Opinião

Conforme descrito no capítulo 2, o processo KDD, em sua essência, consiste na transformação de dados em conhecimento e, para isso, faz-se necessária a utilização de técnicas de mineração de dados que estão divididas em etapas.

Este trabalho tem como objetivo a criação de um modelo para predição da bolsa de valores baseado na mineração de opinião, em que se utilizará como estudo de caso a Empresa Petrobras (PETR4)[1], esta foi escolhida por ser uma empresa brasileira, de grande porte e já consolidada no mercado. Para alcançar referido fim, será necessário usar o processo KDD, que fará parte desta abordagem. Importante ressaltar que este estudo não tem a pretensão de ser a principal ferramenta na tomada de decisão dentro do domínio, e sim, fornecer informações, que quando utilizadas em conjunto com outras técnicas ou ferramentas, venham a contribuir no momento de comprar ou vender um ativo.

O modelo proposto neste trabalho faz parte de uma linha de pesquisa que vem sendo desenvolvida no Laboratório de Sistemas Inteligentes (LSI), da Universidade Federal do Maranhão, voltada para a predição da bolsa de valores. Em citado projeto já foram publicados vários trabalhos que fizeram uso da Inteligência Artificial para este tipo de predição, dentre os quais, merecem destaque de Almeida (2015), em seu trabalho intitulado "*Modelo de Predição para*

[1]Petróleo Brasileiro S.A. é uma empresa de capital aberto, cujo acionista majoritário é o Governo do Brasil. É, portanto, uma empresa estatal de economia mista.

o Mercado Acionário Baseado na Lógica Fuzzy", e o de Nascimento (2015), com "*Um Serviço Baseado em Algoritmos Genéticos para Predição da Bolsa de Valores*".

Vale acentuar que cada um dos trabalhos citados tem características diferentes sobre a predição da bolsa de valores, fortalecendo a ideia do projeto com múltiplas abordagens, o que facilita, sob o ponto de vista do investidor ou negociador, uma visão diferenciada sobre várias percepções do mesmo mercado.

O intuito do presente trabalho, dentro do contexto da plataforma do LSI, é dar uma nova perspectiva de análise para o investidor, agora sob o ponto de vista da mineração de opinião, obtidos através de dados oriundos das postagens no *Twitter*.

4.1 Metodologia

Tendo em vista a dificuldade em prever o comportamento das ações na bolsa de valores, este trabalho propõe um modelo que utiliza as fases do processo KDD (Seleção dos dados, Pré-processamento, Transformação, Mineração de dados, Avaliação e Interpretação dos Resultados) para a extração do conhecimento, visando a predição da tendência dos ativos de uma determinada empresa, neste caso específico, a Petrobrás. A Figura 4.1 ilustra um diagrama de blocos contendo a metodologia adotada para este trabalho.

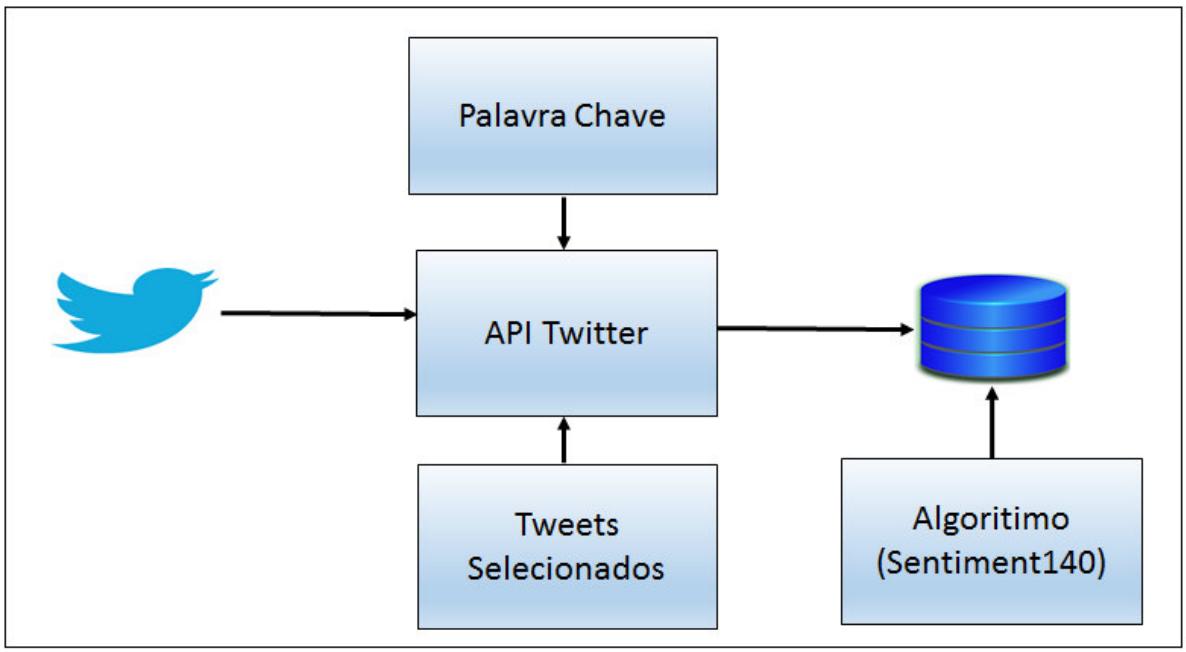


Figura 4.1: Visão Geral da Metodologia(Fonte:Autor)

A Figura acima mostra o fluxo das etapas empregadas na metodologia para a concepção do modelo, que consistem na requisição de busca no *Twitter* via API, mediante a escolha de uma palavra-chave. Esta API, que é própria do *Twitter*, por sua vez, retorna um conjunto de *tweets* que satisfizeram aos critérios de pesquisa preestabelecidos. Dando continuidade ao fluxo das etapas, o resultado da pesquisa é armazenado em um banco de dados, originando o *corpus* que, em seguida, já em outra etapa, passa por um pré-processamento ou limpeza para que finalmente seja submetido ao algoritmo que fará a extração do sentimento.

4.1.1 Coleta de Dados do *Twitter*

O ambiente escolhido para a coleta dos dados textuais necessários para a consecução desta pesquisa foi o *Twitter*. A opção estratégica por essa rede social se deve em grande parte a características únicas que ela apresenta quando em comparação com outras redes (vide seção 2.2.2, sobre o *Twitter*).

Importante pontuar que, atualmente, o *microblog* tem se consolidado como fonte de dados para diversas pesquisas nas mais diversas áreas, como ciências, comércio, indústria. A facilidade no seu acesso, que consiste em estar disponível em várias plataformas, faz com que os usuários interajam em curtos intervalos de tempo, tendo como resultado um enorme legado de dados, normalmente ricos em comentários sobre os mais diversificados assuntos, como mostra a Figura 4.2, que toma como referência o que é "twittado" no Brasil.

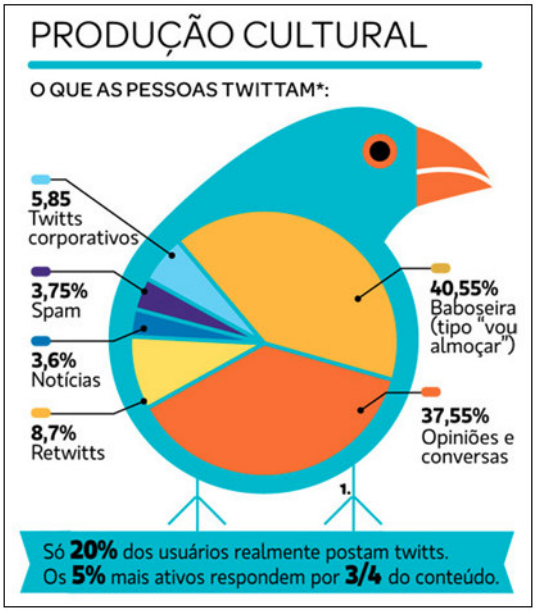


Figura 4.2: *Tweets* por assunto Super (2010)

É nessa grande quantidade de dados não estruturados, que fazem referência a diversos assuntos, que inúmeros pesquisadores se debruçam e buscam extrair conhecimento.

4.1.2 Construção do *Corpus*

As mensagens que serão coletadas advêm de publicações feitas pelos usuários devidamente cadastrados no *Twitter* e cujos *tweets* atendam aos critérios de consulta preestabelecidos através de uma palavra-chave, idioma, localização e período. Neste caso específico, será utilizada a palavra "Petrobras" como chave de pesquisa, cujos *tweets* estejam no idioma português, originários do Brasil referenciando a empresa que será alvo do estudo.

Com o propósito de ter acesso às publicações foi desenvolvida uma aplicação em PHP [2] (vide anexo A), que em conjunto com a API do *Twitter*, disponibilizada pelo próprio site, responsabiliza-se por selecionar e armazenar esses dados. A conexão com o *Twitter* via API trata-se de um recurso que, mediante credenciamento do usuário como desenvolvedor (vide anexo B), permite acesso às mensagens através do protocolo OAUTH [3].

Dentre as etapas que envolvem a coleta de dados, a escolha da palavra-chave adequada a ser consultada é de extrema importância para o resultado da pesquisa, visto que ele é o principal critério para a busca do que está sendo publicado sob a forma de mensagens. Como resultado da extração dos dados, foram obtidos cerca de 3.000 *tweets*/dia, totalizando aproximadamente 40.000 mensagens entre 01/09/2015 e 20/11/2015, no horário de 10h00min a 17h00min (horário de Brasília), lapso em que o Mercado estava em plena operação. A Figura 4.3 ilustra uma sequência de *tweets* selecionados pela API quando submetidos aos critérios de pesquisa adotados.

[2] PHP é uma linguagem de script de propósito geral popular que é especialmente adequado para o desenvolvimento web. Disponível em http://php.net

[3] Um protocolo de autenticação aberto utilizado pela maioria das redes sociais e e-mails. Este framework permite que aplicações terceiras obtenham acesso limitado à um serviço HTTP. Mais informações em: http://oauth.net/

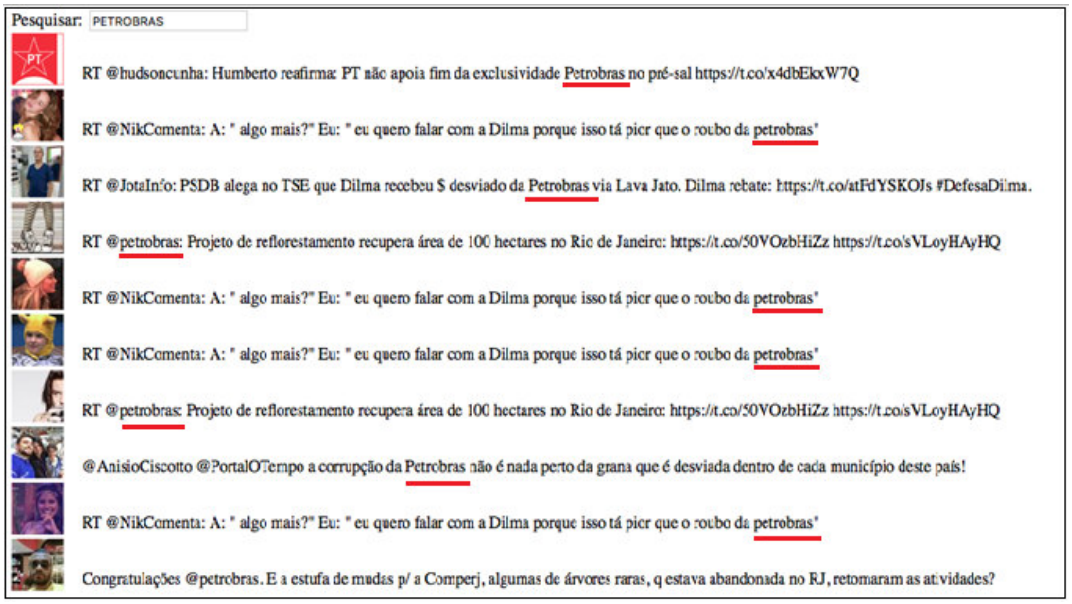


Figura 4.3: Exemplo de *tweets* selecionados com o critério de pesquisa adotado.

Os dados que satisfizeram aos critérios de consulta foram extraídos e armazenados em um banco de dados *MYSQL* [4], como mostra a Figura 4.4.

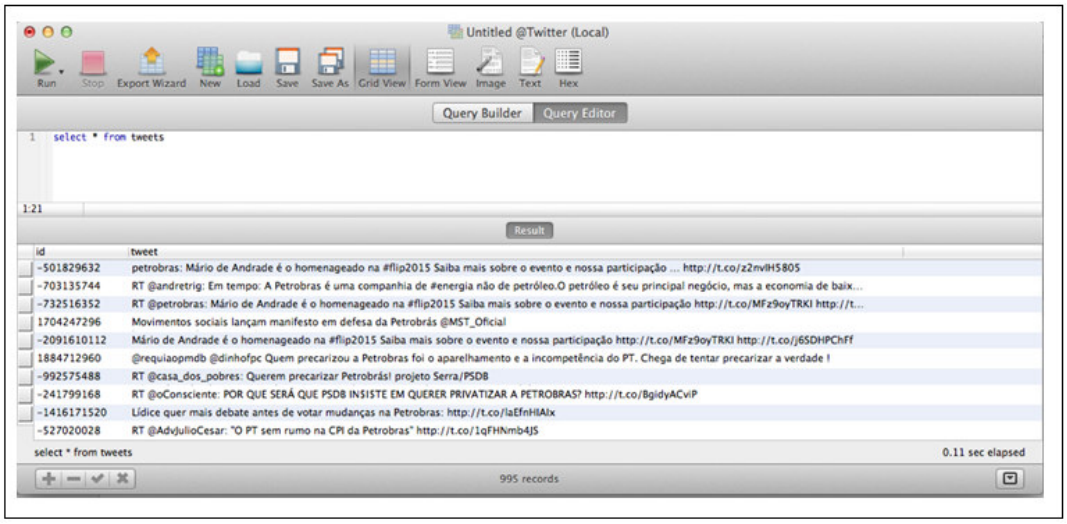


Figura 4.4: Tweets Selecionados e Armazenados em Banco de Dados

4.1.3 Pré-Processamento ou Limpeza dos Dados

As mensagens trocadas nas redes sociais, e com o *Twitter* não é diferente, não obedecem de fato a nenhuma regra gramatical; em geral são textos livres, não estruturados e cheios dos mais

[4] É um sistema de gerenciamento de banco de dados. Mais informações em https://www.mysql.com

diversos termos que dificultam o PLN, tratando-se de gírias, neologismos, ironias, dentre outras expressões de difícil entendimento por parte dos algoritmos que se destinam a essa finalidade. Nesse contexto, os dados precisam passar por uma fase de pré-processamento ou limpeza (etapa que é parte integrante do processo KDD), de forma que fiquem o mais normalizados possível, facilitando, assim, o trabalho do algoritmo responsável pela mineração de opinião. Procedimentos para suprimir caracteres e cadeia de caracteres desnecessários também foram adotados nesta etapa, relacionados a baixo:

- **Acentos** - Todos os acentos foram removidos das palavras;
- **Caracteres especiais** - Tais como: *&%#-;
- **Retwittes** [5] - O "RT"no início das publicações foi removido;
- **Links/Url** - Foram removidos os *links/url's* encontrados nas publicações.

Ao final desta etapa houve uma reestruturação no quantitativo dos dados previamente coletados, ocasionando uma redução de aproximadamente 31% dos *tweets*, os quais não satisfizeram aos requisitos durante o pré-processamento, como, por exemplo, os *retwittes*, que se fossem considerados ponderariam a amostra de forma equivocada, interferindo no resultado da pesquisa.

Convém ressaltar que as mensagens foram coletadas originalmente no idioma português, conforme critério de pesquisa previamente definido, tendo em vista preservar aqueles comentários que, dentre outros fatores, possam sofrer influência decorrente do aspecto geográfico, por estarem no mesmo território que a Petrobras. Outro fator determinante na escolha do idioma de consulta foi a pequena quantidade de amostras resultantes quando o idioma escolhido foi o inglês. No entanto, para que o *corpus* se adequasse ao algoritmo de classificação do sentimento, cujo dicionário léxico orginalmente é na língua inglesa, foi necessário introduzir uma nova etapa no tratamento dos dados, que consistiu em traduzir os textos para o idioma inglês. Para esta tarefa, utilizou-se o *Google Translate*[6].

Uma vez efetuado pré-processamento no *corpus*, os dados estão prontos para uma nova etapa, que será demonstrada na seção seguinte, onde serão submetidos ao algoritmo de classificação da polaridade do sentimento.

[5] O retweet é usado por usuários que querem divulgar uma publicação feita por outras pessoas em seu perfil pessoal

[6] É um serviço gratuito de tradução de textos disponibilizado pela empresa *Google*. Disponível em: https://translate.google.com.br/?hl=pt-br

4.1.4 Mineração da Opinião

Nesta seção será iniciada a mineração da opinião propriamente dita, sendo que, para esta finalidade, serão utilizadas as etapas de transformação e mineração dos dados, também descritas no processo de descoberta do conhecimento e detalhadas no capítulo 2 deste trabalho.

4.1.4.1 Algoritmo Sentiment140

Inicialmente foram pré-selecionadas ferramentas que são bastante utilizadas para mineração de opinião em diversas áreas. Após terem sido submetidas a um mesmo teste, e utilizando a mesma amostra, a opção escolhida foi o léxico Sentiment140 Mohammad et al. (2013), por ter se revelado mais eficiente na classificação das mensagens com as características do *Twitter*. A Tabela 4.1 mostra as ferramentas utilizadas na fase de testes, de modo a encontrar a que melhor se adequasse a este trabalho, assim como suas características.

Tabela 4.1: Algumas ferramentas para mineração de opinião

Ferramenta	**Características**
SentiWordNet	Dicionário lexico e classificação do sentimento obtidos por aprendizagem de máquina
SenticNet	A abordagem baseada em PLN para inferir a polaridade no nível semântico
Sentiment140	API que permite a classificação de *tweets* para polaridades positivo, negativo e neutro

Na fase de testes, que culminou com a escolha do Sentiment140, primeiramente foram selecionados 100 *tweets* de forma aleatória, pertencentes ao *corpus* que totaliza 55.000 registros, para que fossem classificados manualmente e posteriormente comparados com os resultados obtidos nas ferramentas. Na Tabela 4.2 ficam evidentes os melhores resultados quando as mensagens foram submetidas a classificação pelo léxico Sentiment140, isto considerando-se a classificação feita manualmente na amostragem.

Tabela 4.2: Índice de acerto na classificação dos *tweets*

Class.Manual	**Senticnet**	**SentiWordNet**	**Sentiment140**
100%	55%	52%	**65%**

O Sentiment140 é um léxico [7] concebido especificamente para identificar o sentimento contido em uma publicação do *Twitter*, podendo ser utilizado através de API [8].

[7] É uma lista predefinida de palavras, onde cada palavra está associada com um sentimento específico
[8] Disponível em: //www.sentiment140.com

A concepção do *corpus* deste léxico foi estruturada de forma automática (não supervisionada) a partir de uma coleção de 1,6 milhão de *tweets* compostos por *emoticons*[9] positivos e negativos. Nele os *tweets* são rotulados em positivo ou negativo, de acordo com o respectivo *emoticon*. A partir da rotulação automática, verificou-se quais palavras ocorriam com maior frequência em *tweets* positivos ou negativos, dando origem a um dicionário com mais de 1 milhão de termos, distribuídos em 62.468 unigramas[10], 677.698 e 480.010 bigramas pares. A classificação do sentimento de um *tweet* "w" é calculado através do valor do seu score, como demonstrado na fórmula abaixo:

$$score(w) = PMI(w, positive) - PMI(w, negative) \tag{4.1}$$

Nesse contexto, PMI (*Pointwise Mutual Information*) receberá as ocorrências predefinidas como positivas e negativas da expressão, respectivamente. Uma pontuação positiva indica associação com o sentimento positivo, enquanto uma pontuação negativa indica associação com o sentimento negativo e uma pontuação neutra indica ausência de sentimento, como ilustrado na Figura 4.5. Na frase constante em tal figura, a palavra "decadente"[11] é o termo referenciado no dicionário léxico do algoritmo, que é decisivo na classificação do sentimento, por estar identificado como negativo.

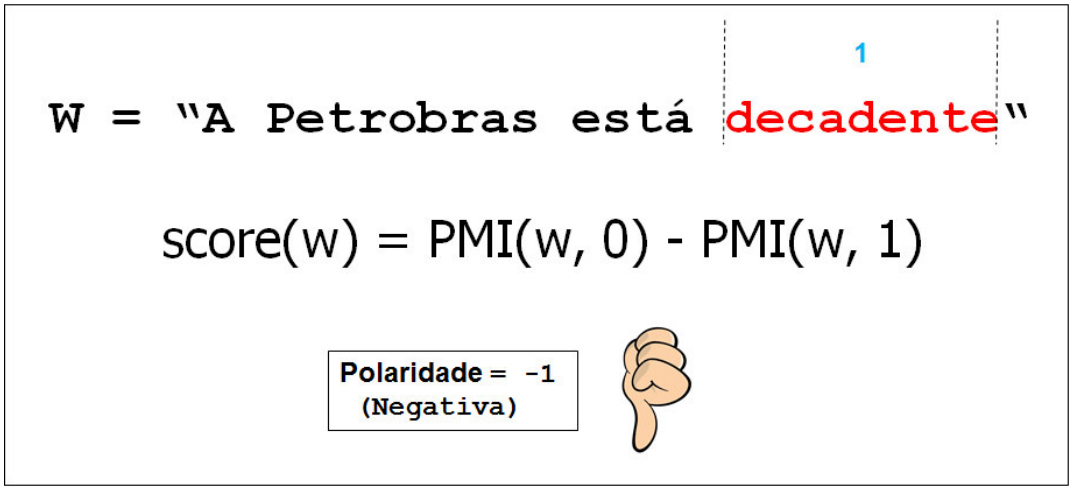


Figura 4.5: Representação da mineração de opinião de uma sentença pelo Sentiment140

[9] Um emoticon é um tipo de expressão facial usado em comunicação escrita on-line, em mensagens de texto e em clientes de mensagem instantânea. Geralmente utilizada para expressar o humor atual do autor da mensagem.

[10] N-Grama é uma sequencia de ítens dentro de uma frase, podendo ser palavras, letras, símbolos etc. Um n-grama de tamanho 1 é chamado de unigrama, de tamanho de 2, de bigrama, de tamanho 3 é chamado de trigrama, de 4 em diante é n-grama.

[11] A frase foi traduzida para o português para fins didáticos, a tradução da palavra em destaque para o inglês é: "decadent", como encontrada no dicionário léxico do algoritmo.

A exemplo da maior parte dos métodos e técnicas disponíveis para mineração de opinião, seu conteúdo se encontra disponível na língua inglesa, podendo ser redefinido através de parâmetro para o espanhol.

4.1.4.2 Mineração de Opinião

A mineração de opinião utilizada nesta pesquisa está subdividida em duas etapas, a saber: na primeira será analisado individualmente cada *tweet*, agrupando-os por dia, em busca do sentimento individual contido em cada postagem; em seguida será encontrado o sentimento coletivo diário ($S_{c(d)}$), demonstrado a partir da equação 4.2, como sendo a diferença do somatório das postagens positivas menos as postagens negativas do dia. A relação da pontuação resultante da operação com o sentimento segue o mesmo modelo da equação 4.1.

$$S_{\mathrm{c(d)}} = \sum_{Positivos} - \sum_{Negativos} \tag{4.2}$$

A fim de obter o sentimento coletivo diário do que é comentado no *Twitter* sobre o domínio, foram selecionados de forma aleatória 1.000 *tweets*/dia do *corpus*, sendo que, após as etapas de pré-processamento e transformação dos dados, o resultado encontrado foi disposto conforme a Figura 4.6, que mostra um fragmento do período em análise (08/10/2015 a 26/10/2015), onde se observa uma predominância do humor "negativo" sobre o "positivo", este ultimo obteve apenas duas ocorrências, uma no dia 09/10/2015 e outra no aos 26/10/2015.

DIA	MENSAGENS (Tweet's)			
	Positivo	Neutro	Negativo	Sentimento
2015-10-08	366	99	**535**	
2015-10-09	**499**	74	425	
2015-10-12	355	88	**443**	
2015-10-13	408	121	**471**	
2015-10-14	344	103	**553**	
2015-10-15	398	112	**602**	
2015-10-16	323	96	**581**	
2015-10-19	313	80	**607**	
2015-10-20	406	79	**515**	
2015-10-21	248	63	**687**	
2015-10-22	290	123	**587**	
2015-10-23	403	149	**448**	
2015-10-26	**521**	93	386	

Figura 4.6: Sentimento Coletivo Diário (fragmento da amostra)

Os dados contendo o sentimento coletivo foram dispostos em uma série histórica para posteriormente serem utilizados, juntamente com outros indicadores provenientes da bolsa de valores, o que será mostrado com maiores detalhes no capítulo seguinte.

4.1.5 Coleta de Dados do Mercado

Inicialmente, os atributos escolhidos para o treinamento e testes foram coletados a partir de uma série histórica do IBOVESPA referente à ação PETR4, dentro do período estudado, contendo: preço de abertura, preço mínimo, preço máximo, preço de fechamento e situação de fechamento (alta/baixa). A escolha por estes indicadores foi motivada pela sua capacidade de expressar, em números, a situação do mercado em relação a determinado ativo, desde a abertura até o fechamento de um dia de negociação na bolsa.

A série histórica contendo os indicadores foi obtida via requisição *web*, através do *Yahoo Finance*[12]. O processo de solicitação das informações acontece diretamente no *browser*[13], através de uma linha de comando parametrizada contento informações como código da ação e período a ser analisado. O resultado da requisição no formato CSV está disponível no anexo C.

4.1.6 Classificador (SVM)

A utilização de um classificador para o alcance do objetivo deste trabalho é de grande importância, sobretudo na etapa que envolve a descoberta do conhecimento. Para tal tarefa, será adotado o SVM. Entretanto, antes que sua escolha fosse definida, outro algoritmo foi testado, o *Naive Bayes*, que não obteve resultados tão bons quanto o SVM, considerando o mesmo conjunto de dados contendo indicadores do mercado financeiro.

Os algoritmos foram escolhidos inicialmente por serem largamente utilizados em soluções semelhantes à que ora buscam-se, onde se combinam mineração de opinião e predição. A Tabela 4.3 mostra os índices de acertos para a amostra, utilizando os algoritmos mencionados, sendo evidente o melhor resultado proporcionado pelo SVM, o que justifica sua escolha.

[12]Serviço disponibilizado pelo Yahoo!, que fornece informações financeiras, notícias, dados e comentários, incluindo cotações de ações , comunicados de imprensa , relatórios financeiros. Disponível em: http://finance.yahoo.com

[13]É um programa desenvolvido para permitir a navegação do usuário pela *web*.

Tabela 4.3: Comparação entre *Naive Bayes* e SVM

Algorítmo	**Instâncias corretamente classificadas**
Naive Bayes	35.29%
SVM	**47.05%**

Convém ressaltar que a utilização de algoritmos de aprendizagem de máquina não é o foco da pesquisa, portanto, o aprofundamento em cálculos e deduções meramente matemáticas será abstraído, limitando-se à utilização de seus algoritmos de forma encapsulada para fins de atender a pesquisa.

4.1.7 Inserção do "Sentimento" ao Modelo

Um dos objetivos da pesquisa consiste em demonstrar o quanto o sentimento pode ser eficiente quando inserido em um cenário de predição da bolsa de valores. Logo, a fim de que se tenha um parâmetro de análise, será necessário dividir a tarefa de mineração dos dados em dois momentos, ambos com as mesmas características de aprendizado, testes e algoritmo. Tem-se como propósito estabelecer uma comparação entre dois cenários, um apenas com os indicadores financeiros e o outro com a inserção de um novo atributo na relação, representado pelo sentimento coletivo.

Nos experimentos, em busca de prever a classe, aqui definida como "comportamento", assim denominada por fazer referência ao comportamento em relação ao índice de fechamento quando comparado com o dia anterior. Podendo esta, assumir dois valores distintos: "alta" ou "baixa". Neste intuito foi utilizado um modelo composto por um conjunto de dados previamente rotulados, a fim de efetuar o treinamento supervisionado do classificador, utilizando o algoritmo SVM.

Os dados foram divididos em dois grupos, sendo o primeiro para treinamento, em que será utilizado 70% das instâncias, e o segundo, com 30% para testes[14], em um universo de 55 instâncias. A Figura 4.7 ilustra um fragmento do arquivo .arff e a inserção do novo atributo "sentimento" na "Etapa-2".

[14]Esse método é chamado *hold out* e consiste em separar os dados em dois grupos, um chamado grupo de treinamento e outro de grupo de teste.

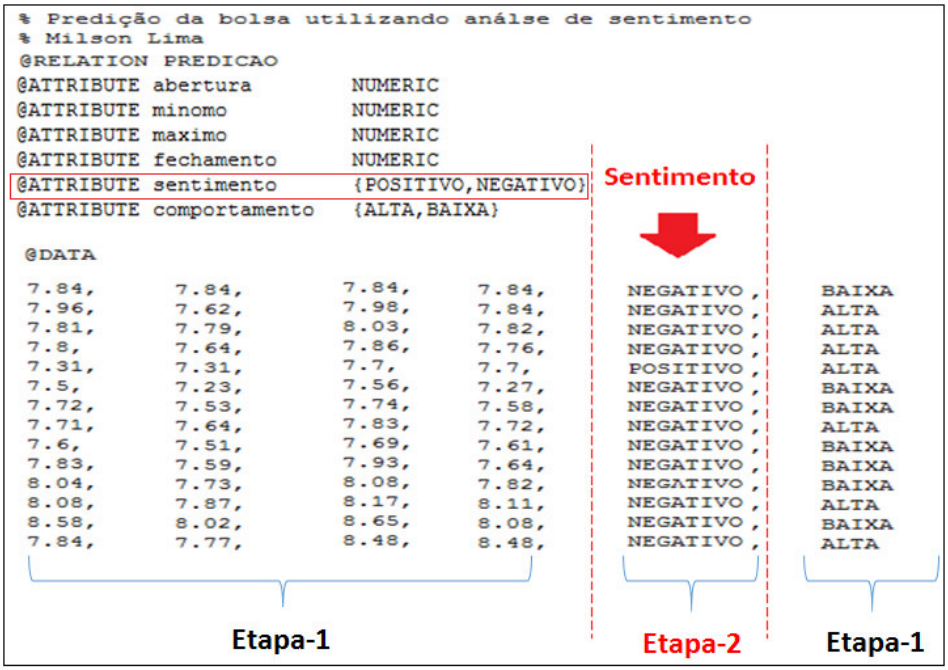


Figura 4.7: Fragmento do arquivo .arff e seu novo atributo “Sentimento”

A Figura 4.8 expõe as duas etapas propostas, assim como o fluxo dos dados, desde a relação de atributos sob a forma de arquivo até a avaliação dos resultados. No princípio, os dois conjuntos de dados são submetidos ao método *Hold Out*, para que sejam treinados e testados pelo classificador sob as condições definidas. Uma vez criado o modelo, seu aprendizado é testado para fins de predição.

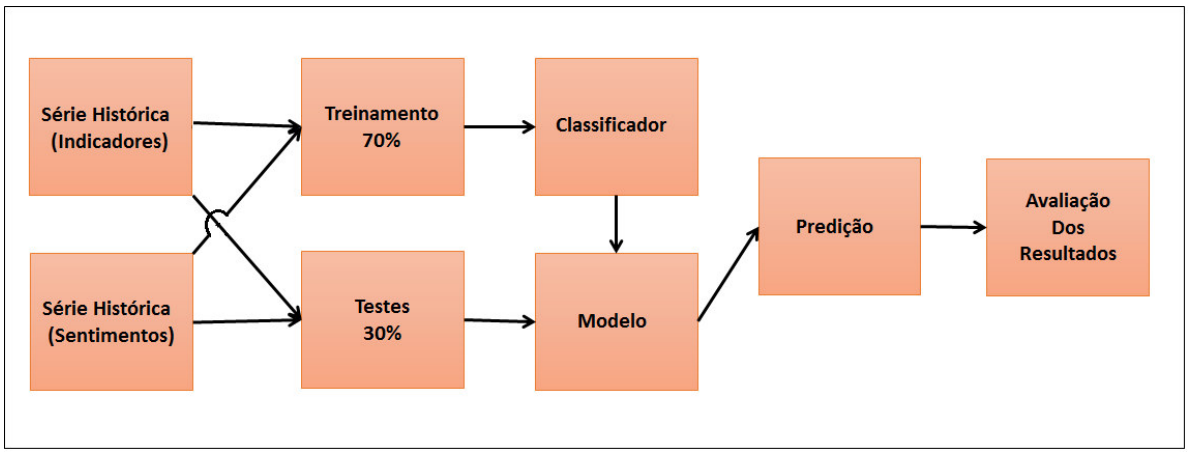


Figura 4.8: Fluxo das duas etapas da mineração de dados (Adaptado do método *hold out*)

Ao término das etapas demonstradas neste capítulo, foi criado um protótipo que será apresentado na seção a seguir, bem como sua modelagem, tecnologias utilizadas para sua concepção e interface.

4.2 Protótipo

Segundo Berkun (2000), por definição, protótipo é qualquer representação da idéia de um produto em projeto. No contexto da Engenharia de Software, protótipos podem ser entendidos como sendo a representação gráfica, não necessariamente funcional, de um sistema em fase de projeto, seja construção ou re-engenharia Rudd et al. (1996).

Sendo assim, visando automatizar as diversas etapas da metodologia descrita neste capítulo, implementou-se um protótipo que faz uso do modelo criado. Como consequência da prototipação, houve o encapsulamento de diversas etapas já demonstradas neste capítulo, o que torna mais intuitiva a utilização da metodologia, auxiliada por uma interface que facilitará sua aplicação para outras empresas da bolsa de valores objetivando predizer o mercado.

De uma forma geral o protótipo encapsulará as tarefas demonstradas na Tabela 4.4, estas foram também relacionadas, para efeitos didáticos, com suas respectivas fases dentro do processo KDD.

Tabela 4.4: Tarefas Realizadas pelo Protótipo

Tarefa do Protótipo	**Correnpondência no processo KDD**
Selecionar os dados no *Twitter*, correspondentes a ação em estudo através de API própria	Seleção
Homogeneizar dos dados, retirando possíveis inconsistências que interfiram no resultado	Pré-Processamento ou Limpeza
Formatar os dados dentro dos padrões aceitos pelo algoritmo de classificação (SVM). Nesta etapa já esta abstraída a polarização das mensagens através *Sentiment140*, bem como a sumarização do sentimento coletivo diário ($S_{c(d)}$).	Transformação
Classificar os dados utilizando com base no modelo.	Mineração de Dados
Predizer a tendência de fechamento da ação	Avaliação

Todas essas etapas que fazem parte do protótipo visam um resultado final que é a obtenção do conhecimento, que por sua vez auxiliará na tomada de decisão dentro do processo de compra e venda de uma ação no mercado.

4.2.1 Tecnologias Utilizadas

O protótipo em questão foi desenvolvido predominantemente na linguagem PHP, com exceção do módulo de acesso ao *framework WEKA* (fragmento do código disponível no Anexo D), implementado em *JAVA*[15]. A coleta e polarização dos dados foram feitas por duas API's:

[15] Java é uma linguagem de programação interpretada orientada a objetos desenvolvida na década de 90 por uma equipe de programadores chefiada por James Gosling, na empresa Sun Microsystems

uma disponibilizada pelo próprio *Twitter*, utilizada para a coleta dos dados e a outra o *Sentiment140* para a polarização do sentimento dos *Tweets*. Nos anexos B e E, respectivamente, consta parte do código utilizado para estas finalidades.

Quanto ao armazenamento dos dados coletados e processados, utilizou-se o MYSQL (estrutura disponível no Anexo F). E por fim, como ambiente de programação para concepção do protótipo foi utilizado o NetBeans na versão 7.4[16].

A combinação de todas essas ferramentas que incluem linguagens de programação, banco de dados e ambiente de desenvolvimento, resultaram na concepção do protótipo que será modelado na próxima seção.

4.2.2 Modelagem

Segundo Booch et al. (1996), "um diagrama é uma representação gráfica de um conjunto de elementos, geralmente representados como gráfico conectado de vértices (itens) e arcos (relacionamentos)". A UML (*Unified Modeling Language*) oferece vários diagramas na intenção de facilitar a compreensão e visualização de um sistema sob diversas perspectivas.

Para a construção do protótipo proposto, primeiramente realizou-se a etapa de modelagem UML básica, objetivando uma melhor compreensão do seu objetivo final. Para que isso se concretizasse tornou-se necessário a construção dos diagramas de casos de uso, classe, sequência e atividades de forma a viabilizar a estruturação do mesmo.

4.2.2.1 Diagrama de Caso de Uso

A Figura 4.9, mostra o diagrama de caso de uso do protótipo onde se observa a presença de um único ator, aqui denominado de "Investidor". Responsável por iniciar o processo onde será definido o sentimento coletivo do mercado, coletar a cotação da ação, aplicar o algoritmo SVM até a predição propriamente dita.

[16] O *NetBeans IDE* é um ambiente de desenvolvimento integrado, gratuito e de código aberto para desenvolvedores de *softwares* em diversas linguagens

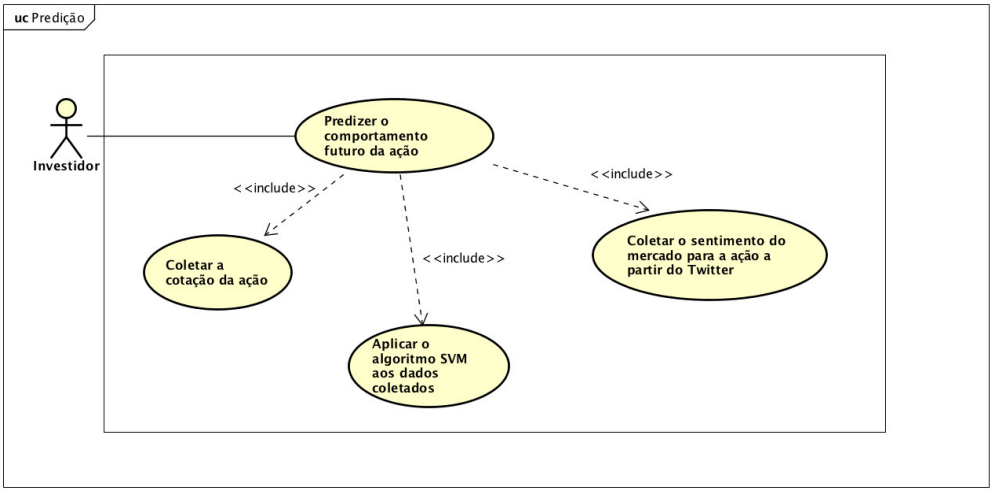


Figura 4.9: Diagrama de Caso de Uso

4.2.2.2 Diagrama de Classes

"O diagrama de Classes está no núcleo do processo de modelagem de objetos. Ele modela as definições de recursos essenciais à operação correta do sistema" Pender (2004).

Na Figura 4.10, onde é mostrado o diagrama de classes, ficam evidenciadas as classes utilizadas pelo protótipo, atributos e relacionamentos. Este último foi utilizado entre as duas classes "tweet" e "corpus" para garantir a consistência dos dados, visto que o "corpus" é formado a partir de dados tratados (pré-processados) provenientes de "tweet".

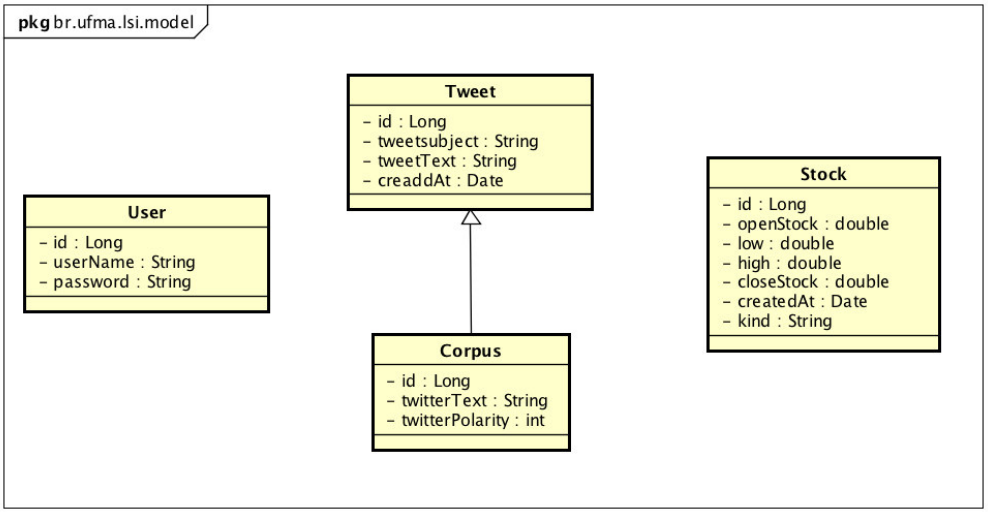


Figura 4.10: Diagrama de Classes

4.2.2.3 Diagrama de Sequência

No diagrama de sequência o objetivo é demonstrar a sequência de mensagens trocadas entre os objetos. Seu objetivo é identificar as interações entre os objetos no decorrer do tempo. Segundo Pender (2004), "O diagrama de sequência utiliza uma visualização orientada para o tempo. Ele usa um conjunto de ícones de objeto e linhas de tempo associadas, chamadas linhas de tempo do objeto, para cada objeto".

No diagrama de sequência da Figura 4.11, estão ilustradas as trocas de mensagens entre objetos e entre atores e objetos. Para este caso específico que retrata a sequência utilizada pelo protótipo, interagem em momentos específicos o ator, aqui denominado "Investidor", o "Controlador" que é responsável por centralizar as informações processadas e redistribui-las de forma sequencialmente lógica entre o "Twitter", "Sentiment140", "Yahoo Finance" e o algoritmo "SVM", para que finalmente possa retornar ao "Investidor" a predição da tendência de fechamento do mercado sobre a ação selecionada.

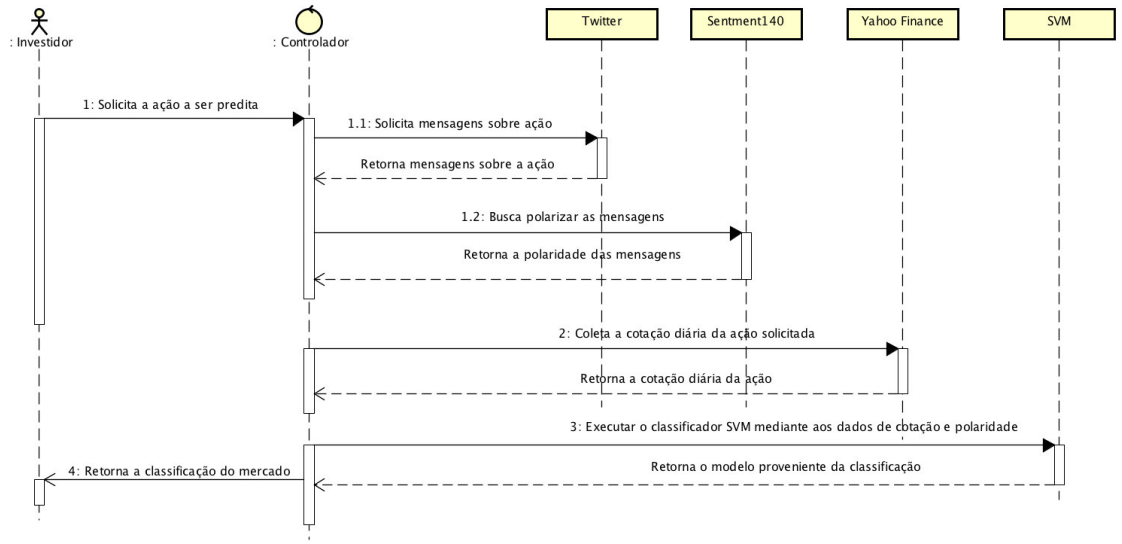


Figura 4.11: Diagrama de Sequência

4.2.2.4 Diagrama de Atividades

O Diagrama de Atividades é utilizado para demonstrar a lógica de programação e regra de negócios. Este diagrama determina as regras essenciais de sequência que se devem ser seguidas para a execução do processo. A Figura 4.12 corresponde ao diagrama de atividades do protótipo, onde foram mapeadas cinco atividades básicas, que vão desde o processamento da palavra-chave que identifica a ação no mercado até a classificação do comportamento futuro

da ação (predição).

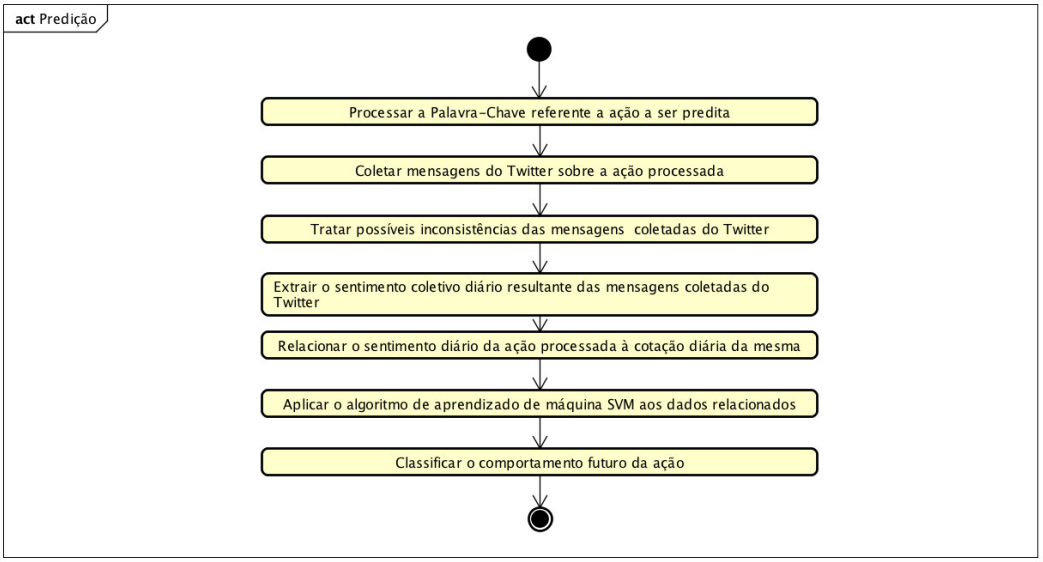


Figura 4.12: Diagrama de Atividades

4.2.3 Interface

Tendo em vista uma melhor visualização dos resultados da pesquisa, bem como automatizar alguns de seus processos tornando-os simples e amigáveis para o usuário final, o protótipo demonstrado nessa seção é composto por algumas telas (páginas), onde é possível analisar resultados através de gráficos e tabelas, ambos construídos de maneira dinâmica. A seguir serão apresentadas algumas interfaces e suas respectivas finalidades.

4.2.3.1 Tela de Login/Logoff

Com o objetivo de permitir o acesso somente a usuários devidamente cadastrados, foi criado um formulário de acesso ao protótipo, demonstrado na Figura 4.13, o mesmo é composto basicamente por dois campos, sendo o primeiro para que seja inserido o nome do usuário e o segundo no formato *passwordchar*, onde os caracteres digitados são substituídos por um asterísco (*), no intuito de proteger os dados digitados.

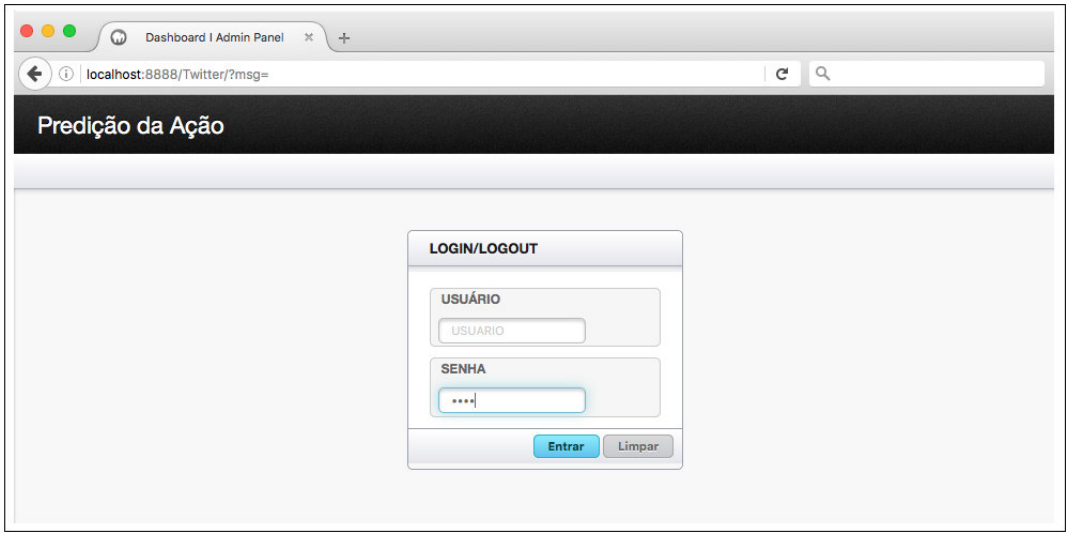


Figura 4.13: Tela de Login/Logoff

4.2.3.2 Tela de Coleta e Classificação dos *Tweets*

A Figura 4.14 mostra a tela responsável pela coleta e classificação dos dados. É neste momento onde é inserida a palavra-chave que servirá de argumento de pesquisa no *Twitter* via API própria.

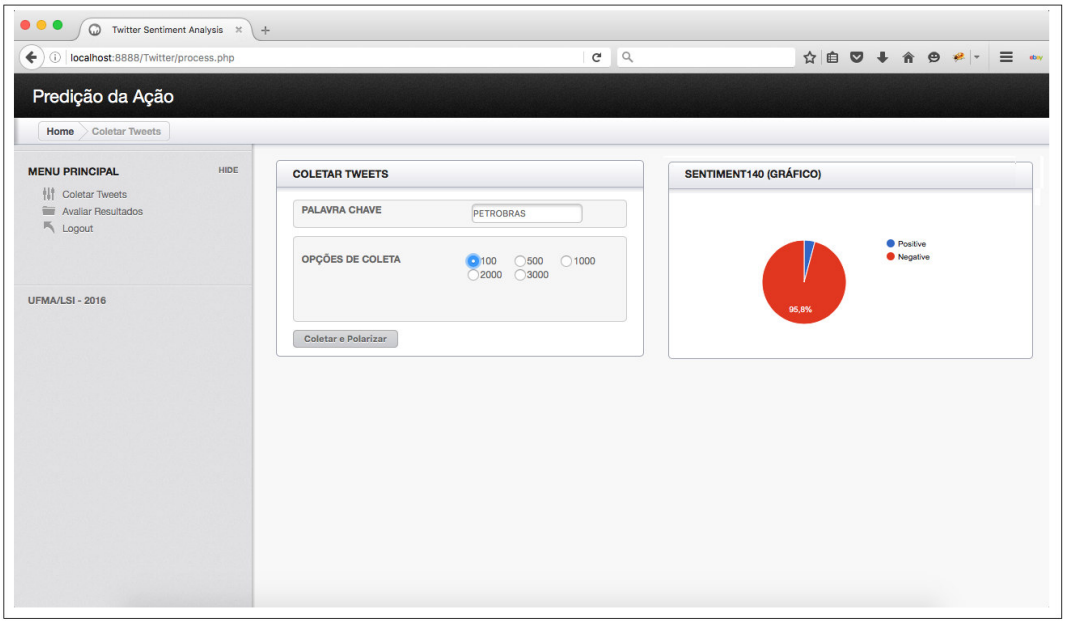


Figura 4.14: Tela para Coleta e Classificação dos *Tweets*

Em um segundo momento, já de posse dos *Tweets* em sua forma original (sem os devidos tratamentos) e devidamente armazenados na tabela denominada “tweets” do banco de dados

MYSQL, os dados são submetidos a um processo de “limpeza” para padroniza-los. Esta etapa é correspondente no processo KDD a etapa de pré-processamento.

Uma vez que os dados foram homogeneizados, eles popularão uma tabela chamada “corpus”, responsável por armazenar o conteúdo do mesmo nome, contendo os dados que serão submetidos ao algoritmo de classificação do sentimento, o *Sentiment140*, retornando para cada *Tweet* sua respectiva polaridade do sentimento (positivo ou negativo).

Ao fim do processamento, será exibido um gráfico demonstrando o resultado em forma de pizza, onde cada fatia representa uma polaridade e a maior dentre elas será o humor coletivo.

No protótipo, esta etapa é a que consome maior tempo para executar suas tarefas, visto que tanto a coleta dos *Tweets*, quanto a classificação do sentimento utilizam recursos de API’s, sendo assim dependem de alguns fatores que podem ocasionar lentidão ao processo, como por exemplo uma baixa velocidade de acesso a *Internet*.

4.2.3.3 Tela de Resultado da Predição

Por fim, o protótipo apresenta sua etapa responsável pela predição propriamente dita, como mostra a Figura 4.15, nesta etapa é estabelecido um relacionamento, entre duas tabelas, uma delas já mencionada (“corpus”) e uma nova no contexto, que aqui foi denominada de “stoks”, tendo por finalidade armazenar os dados das cotações provenientes do *Yahoofinance* sobre a ação em análise informada no campo “Código da Ação”.

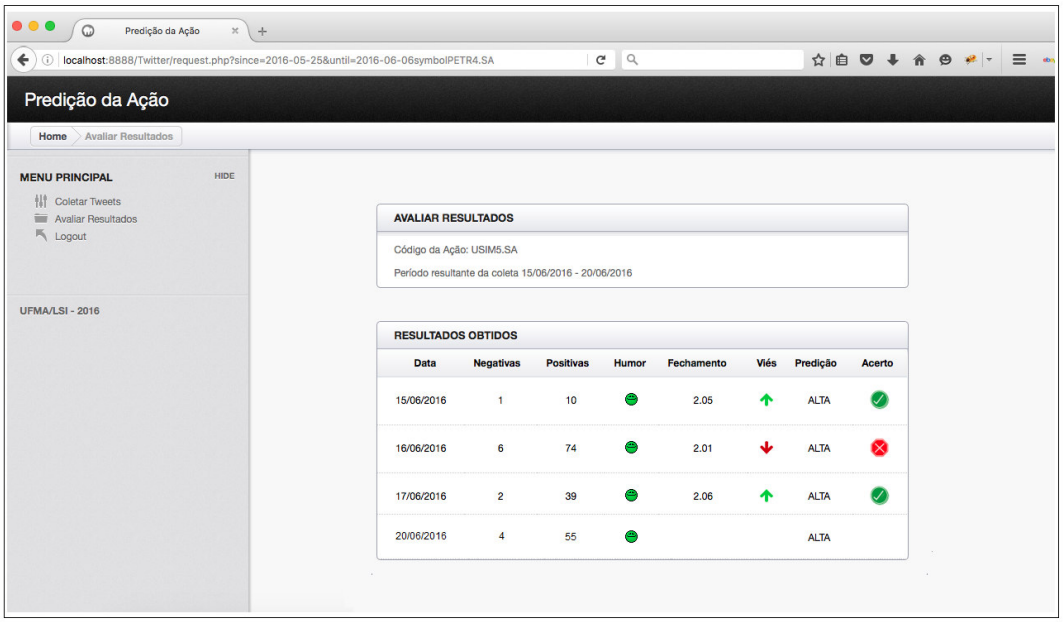


Figura 4.15: Tela de Resultado da Predição

O referido relacionamento tem por objetivo encontrar correspondências entre os dias onde há cotações no *Yahoofinance* e mensagens no *Twitter*, visto que para alguns dias (finais de semana e feriados), existem *Tweets*, no entanto o mercado financeiro esteve sem operar. Para esses casos, as informações coletadas são desprezadas, por não preencherem os parâmetros necessários para a execução do modelo classificador.

Para os demais casos onde há correspondência entre os dados de cotação e o *Twitter* o protótipo exibe uma tabela com as informações já processadas, onde o campo denominado “Predição” representa a classe que contém a informação desejada, ou seja, a predição (ALTA/BAIXA) do fechamento da bolsa para a referida ação.

Convém ressaltar que para a análise de dias onde o mercado financeiro ainda está em plena operação, ou seja, não "fechado", a predição é representada pelo sentimento coletivo predominante até o momento, indicando a tendência de fechamento "alta"quando positivo e "baixa"quando negativo.

4.3 Síntese

Neste capítulo, as etapas e técnicas necessárias para construção do modelo foram apresentadas, onde foi inserida a mineração de opinião dentro do contexto preditivo desse domínio que atualmente é dominado pelas técnicas que fazem uso dados estatísticos associados aos indicadores financeiros. Por fim, foi desenvolvido um protótipo da metodologia utilizada que faz uso do modelo previamente definido e uma modelagem do mesmo.

No próximo capítulo serão analisados os resultados obtidos com a abordagem proposta na pesquisa, validando a relação existente entre o sentimento coletivo extraído das postagens do *Twitter* e o mercado financeiro, quantificando seu nível de acerto através das métricas do algoritmo utilizado.

Capítulo 5

Resultados Obtidos

Utilizando a abordagem proposta no capítulo 4, baseada em grande parte no processo KDD, serão apresentados neste capítulo os resultados obtidos com o modelo proposto, assim como comparações e discussões que visam enriquecer e validar o que será exposto neste capítulo.

5.1 Considerações

Tendo em vista demonstrar que o sentimento contido nas publicações do *Twitter* funciona como uma espécie de termômetro do que pode vir a acontecer no mercado financeiro, em especial com os ativos de uma determinada empresa, serão analisados os dados estruturados no capítulo anterior, fazendo uso de algoritmo de aprendizagem de máquina, buscando extrair conhecimento para que seja usado para fins de predição.

Será avaliada a tendência do ativo da Petrobras (PETR4), de alta ou baixa, em comparação com a variação do humor expresso pelos usuários através de suas postagens. Sempre buscando fundamentar, através de resultados e discussões, o caráter preditivo proporcionado pela mineração de opinião no contexto do mercado financeiro.

Dentro do contexto que envolve a análise de sentimento e predição, um dos grandes desafios é encontrar uma maneira eficiente de classificar os dados para análise. O processo de classificação consiste em encontrar, através de aprendizagem de máquina, uma função que expresse, da melhor forma possível, as classes de dados envolvidas no domínio e, com isso,

tornar automático o processo de classificação para novas instâncias, tendo como referência o modelo para o qual foi treinado.

Neste trabalho, a ferramenta utilizada para o processo de descoberta do conhecimento foi WEKA[1] (*Waikato Environment for Knowledge Analysis*), que é um conjunto de algoritmos de aprendizado de máquina para tarefas de mineração de dados.

A opção por tal ferramenta se deve a algumas características próprias que ela possui, dentre as quais se destaca o fato de ser um *software* livre, de fácil utilização, proporcionada por sua interface gráfica bem amigável.

5.2 Mineração dos Dados

A etapa de mineração de dados também faz parte do processo KDD, sendo responsável por buscar e descobrir informações úteis, que geralmente não estão bem claras.

Oportuno sinalizar que, para Berry e Linoff (1997), mineração de dados é a exploração e análise, por meio automático ou semiautomático, de grandes quantidades de dados, tendo como objetivo a descoberta de padrões e regras. Na presente pesquisa, esta fase consiste em uma mineração de dados, do tipo classificação, utilizando aprendizado supervisionado, pois faz uso de amostras pré-classificadas para treinamento do algoritmo, que neste caso é o SVM, através do *WEKA*.

> *"A inteligência nos permite vasculhar nossa memória observando padrões, inventando regras, tendo novas ideias para fazer previsões sobre o futuro. Harrison (1998)"*

Os dados utilizados nesta pesquisa, originários do *Twitter*, após terem sido submetidos às diversas etapas do processo KDD e os demais, coletados a partir de séries históricas do mercado financeiro, foram dispostos como mostra a Figura 5.1, composta por parte da amostra, compreendida entre os dias 07/10/2015 e 26/10/2015.

[1] Disponível em http://www.cs.waikato.ac.nz/ml/weka/

Neste fragmento os dados foram agrupados em quatro colunas principais, em que se comparam duas das variáveis mais representativas para a pesquisa, a saber, o sentimento e o mercado.

DIA	MENSAGENS (Tweet's)				FECHAMENTO PETR4.SA			ACERTO
	POSITIVO	NEUTRO	NEGATIVO	HUMOR	FECHAMENTO	VIÉS	HUMOR	
07/10/2015	------	------	------	------	8,47	------	------	------
08/10/2015	366	99	**535**		8,75			
09/10/2015	**499**	74	425		8,80			
12/10/2015	355	88	**443**		8,80			
13/10/2015	408	121	**471**		8,13			
14/10/2015	344	103	**553**		7,96			
15/10/2015	398	112	**602**		8,00			
16/10/2015	323	96	**581**		7,95			
19/10/2015	313	80	**607**		7,95			
20/10/2015	406	79	**515**		8,05			
21/10/2015	248	63	**687**		7,75			
22/10/2015	290	123	**587**		8,02			
23/10/2015	403	149	**448**		7,99			
26/10/2015	**521**	93	386		7,85			

LEGENDA
ERRO
ACERTO
NEUTRO

Figura 5.1: Sentimento do *twitter* X Mercado de Ações

Abaixo uma breve explicação sobre o conteúdo de cada coluna principal:

- **DIA** - Dia em análise;
- **MENSAGENS(*tweets*)** - Resultado da análise dos sentimentos (positivo, neutro e negativo) de todas as mensagens processadas (blocos de 1.000 *tweets* para cada dia) e onde o humor coletivo é representado pela variável que expressa o sentimento de maior valor absoluto;
- **FECHAMENTO(PETR4)** - Estão representados os valores referentes ao fechamento do ativo PETR4 na bolsa de valores, tais como valor de fechamento em R$, viés [2];
- **ACERTO** - Resultado da comparação entre o humor das mensagens do *Twitter* e o

[2]Neste contexto, o resultado (alta/baixa) é obtido quando o valor de fechamento é comparado com o dia anterior, assumindo o valor "alta" quando o valor do dia for maior que o do dia anterior e "baixa" para valores cujo dia atual sejam menor que o dia anterior.

humor do mercado, assumindo a cor verde quando essas duas variáveis tiverem o mesmo valor, caso contrário a cor vermelha será adotada.

Ao analisar os dados dispostos na Figura 5.1 que posteriormente teve sua funcionalidade incorporada ao protótipo, como demonstrado na Figura 4.15, percebe-se que há um predomínio do sentimento negativo nas mensagens do *Twitter* na amostra, embora apenas sendo um fragmento, mas essa tendência também se manteve nos dados em sua totalidade. A mencionada disposição do humor negativo presente nas mensagens revela um reflexo do momento pelo qual a empresa em questão está atravessando, envolvida em recentes episódios de escândalos financeiros.

Outro aspecto importante observado, ainda analisando a Figura 5.1, refere-se à divergência do humor presente nas mensagens quando comparado ao do mercado, constatado em alguns dias da série. Tal dado, por si só, pode não ser conclusivo para indicar uma real divergência entre humor dessas duas variáveis, pois, tomando como exemplo o dia 15/10/2015, que indica humor negativo na análise dos *tweets* e positivo no mercado financeiro para PETR4, no entanto, o viés de alta que determinou o sentimento do mercado como positivo, quando comparado ao dia anterior, é muito sutil, sendo de apenas R$ 0,04 (quatro centavos de real), indicando que o humor contido nas mensagens reflete o sentimento coletivo de fato e que o humor do mercado teve uma leve variação, suficiente para mudar sua classificação, mas ainda deixando bem caracterizada a sua mudança de humor. Sendo assim, torna-se possível concluir que isoladamente os dados podem ter uma interpretação equivocada ou duvidosa, necessitando eventualmente de uma visão mais criteriosa em sua análise.

5.3 Avaliação de Desempenho do Modelo

Após serem submetidos às etapas descritas, os modelos entram na fase de avaliação. Os resultados obtidos serão mostrados nesta seção, após a classificação dada pelo algoritmo. O primeiro conjunto de dados que será avaliado é a matriz de confusão [3]. A Tabela 5.1 mostra duas matrizes, uma para cada modelo classificado. O número de acertos para cada classe se localiza na diagonal principal C(Ci,Ci) da matriz C, e os demais elementos C(Ci,Cj), para i $\neq$ j, representam erros na classificação. Abaixo, a matriz de confusão 5.1 e seus elementos, em que:

[3] É uma matriz quadrada n x n onde as colunas representam as n classes reais de entrada e as n linhas as n classes de saída do classificador e onde cada célula fornece o número de elementos classificados.

- **VP** - Verdadeiro positivo: Elementos positivos corretamente classificados;
- **FP** - Falso positivo: Elementos negativos incorretamente classificados como positivos;
- **FN** - Falso negativo: Elementos positivos incorretamente classificada como negativa;
- **VN** - Verdadeiro negativo: Elementos negativos corretamente classificados.

$$C = \begin{bmatrix} VP & FP \\ FN & VN \end{bmatrix} \tag{5.1}$$

Tabela 5.1: Tabela com as Matrizes de Confusão dos Modelos Após o Treinamento

Matrizes de Confusão			
Modelo	a	b	Classificado como:
Indicadores	4	5	a = alta
	0	8	b = baixa
Sentimentos	5	0	a = alta
	3	8	b = baixa

Sendo assim, percebe-se que, das 17 instâncias, que correspondem a 30% dos dados utilizados para testes, no primeiro modelo, o elemento da matriz C(1,2), aqui representando a quantidade de elementos classificados como falsos positivos (FP), contém 5 instâncias classificadas incorretamente. Ao passo que, considerando a posição equivalente no segundo modelo, onde foi inserido o atributo ”sentimento”, pode-se perceber que houve um ganho significativo na classificação, pois nenhum elemento foi classificado incorretamente.

A matriz de confusão também revela outra informação importante sobre os resultados envolvendo os dois modelos, agora analisando os elementos da coluna M(1,1), da diagonal principal, denominado verdadeiro positivo (VP), que são os valores corretamente classificados como positivos pelo algoritmo. No modelo que utiliza somente os indicadores de mercado, o valor encontrado é menor do que no modelo onde há a presença do sentimento, respectivamente 4 e 5, indicando um melhor resultado na classificação.

Por fim, somando os valores da diagonal principal, C(1,1) e C(2,2), que indica o número de acertos em cada classe, chega-se aos seguintes valores: 13 e 12, respectivamente. Esses valores encontrados também demonstram a superioridade dos resultados onde há a presença do atributo sentimento, aqui representado pelo valor 13. O mesmo raciocínio também pode

ser aplicado na diagonal secundária, C(1,2) e C(2,1), em que estão dispostos os valores classificados incorretamente, onde o modelo resultante dos indicadores de mercado tem 5 amostras classificadas indevidamente quando comparado ao outro modelo com apenas 3 amostras.

Ainda sobre os resultados obtidos, baseados nas métricas derivadas da matriz de confusão, a Tabela 5.2 demostra alguns indicadores de desempenhos, comparando-os entre os modelos.

Tabela 5.2: Indicadores de Desempenho dos Modelos

Métrica	**Fórmula**	**Descrição**	**Valores(%)**	
			Indicadores	**Sentimento**
Acurácia (acertos)	(VP+VN)/Total	Total de Acertos	**70,58**	**81,25**
Erros	(FP+FN)/Total	Total de Erros	29,41	18,75
Precisão (média)	VP/(VP+FP)	Quantidade Real de Predições Positivas	81,90	88,30

Nessa nova análise, a exemplo da anterior, agora utilizando os indicadores acurácia, erros e precisões, percebe-se que o modelo resultante da classificação em que foi inserido o atributo sentimento obteve melhores resultados. Os anexos G e H ilustram a telas de saída do *WEKA* com os resultados das duas etapas.

Ao comparar esta pesquisa, de forma breve, com um dos trabalhos relacionados, intitulado "*A Era de um Mercado Social: A Relação Entre o Twitter e o Mercado Acionista*", referenciado no capítulo 3, podem-se enumerar diferenças, embora os trabalhos tenham domínios iguais, ou seja, a predição para a bolsa de valores.

No que diz respeito à metodologia, uma das diferenças que merece ser destacada concerne à coleta dos dados, eis que, enquanto o trabalho citado definiu o seu universo de coletar as postagens no idioma inglês, a presente pesquisa julgou que o idioma do país onde a empresa fruto da investigação está sediada deve ser o idioma que servirá como um dos critérios de pesquisa durante a coleta dos dados. Isto ocorreu por se entender que há uma quantidade maior de informações importantes nesses conteúdos provenientes em sua grande maioria, de usuários do mesmo país onde a empresa está instalada e, por isso, mais próximos das zonas de onde são gerados os debates e discussões sobre a mesma.

No que tange o processo de classificação da polaridade, a utilização do algoritmo *Sentiment140* nesta pesquisa, foi uma escolha determinante para os resultados alcançados, visto que o mesmo apresentou melhores índices de acertos na classificação, quando comparado a outras ferramentas, uma vez que seu dicionário léxico foi concebido tomando como base o próprio *Twitter*, como já descrito no no capítulo 4. Em contra partida o trabalho Bernardo (2014) fez uso de uma técnica léxica utilizada por Hu e Liu (2004) que recorre a um vetor de palavras

pré-selecionadas e compara as mesmas com o conteúdo do texto para associar o mesmo a sentimentos positivos ou negativos. O autor também relata que foi necessário adicionar *emoticons* e palavras a lista, afim de melhorar seus resultados, o que não foi necessário neste trabalho quando da utilização do *Sentiment140*.

Outro aspecto cuja diferença merece ser destacada diz respeito ao agrupamento dos dados, pois, enquanto o autor do trabalho mencionado testa vários tipos de agrupamentos temporais (horas e dias) com os dados do *Twitter*, concluindo que o seu resultado é diretamente influenciado pela forma como os dados são agrupados, o vertente trabalho agrupa os dados da mesma fonte *Twitter*, de maneira diferente, procurando agrupá-los por dia e por polaridade para depois definir o que aqui foi denominado de sentimento coletivo diário ($S_{c(d)}$). Como já detalhado no capítulo 4.

Ainda comparando os trabalhos, outra diferença bem marcante é o quantitativo de empresas envolvidas. Enquanto o trabalho Bernardo (2014) utiliza como estudo de caso várias empresas (vide cap.3) estrategicamente escolhidas por terem seus ativos negociados em diversos países, principalmente os de língua inglesa, o presente estudo utiliza apenas uma, a Petrobras, muito embora o referido modelo poderá ser aplicado a qualquer empresa que possua ativos na bolsa de valores. Este aspecto, segundo o autor referenciado, é decisivo para seus resultados, interferindo diretamente, pois, segundo ele, há empresas mais sensíveis ao sentimento expresso nas publicações do *Twitter* do que outras. Aspecto que não pode ser constatado neste trabalho, por não possuir número suficiente de empresas investigadas para tal conclusão.

5.4 Testes

Visando validar a eficiência do modelo proposto, o protótipo analisou os dados de outras empresa no mercado de ações, empresas estas com características bem diferentes da até então utilizada neste trabalho, a Petrobras.

A Usiminas [4], foi uma das Empresas utilizada para esta validação. Diferentemente da Petrobras, a Usiminas não é uma empresa que apresenta sua imagem associada ao Estado, visto que foi privatizada no ano de 1991, fato este que foi uma das motivações para sua escolha.

Outra empresa submetida aos testes de validação foi o Banco do Brasil [5], escolhido por ser

[4] Usiminas - Usinas Siderurgicas de Minas Gerais S.A é uma empresa do setor siderúrgico líder na produção e comercialização de aços.

[5] Banco do Brasil S.A. é uma instituição financeira brasileira, constituída na forma de sociedade de economia mista, com participação da União brasileira em 68,7% das ações.

representante do grupo das Empresas financeiras, diferentemente das anteriores.

Ainda no que tange a fase de testes, a Empresa Petrobras foi mais uma vez submetida a esta etapa, estendendo-se além daqueles já utilizados pelo algoritmo durante a fase de concepção do modelo em questão. A Tabela 5.3 demonstra o índice de acertos obtido quando comparado o fechamento real e a predição sugerida pelo modelo em diferentes datas do mês de maio e junho do ano de 2016.

Tabela 5.3: Algumas ferramentas para mineração de opinião

PETROBRAS (PETR4)	**USIMINAS (USIM5)**	**BB (BBAS3)**
66.6%	33.3%	66.6%

Os testes aplicados, agora em outras empresas de setores diferentes da economia, apontam que os índices de acertos foram inferiores a acurácia do modelo constatada durante a fase de treino e testes do modelo.

5.5 Síntese

Neste capítulo foi demonstrado, através da abordagem proposta, um modelo baseado em mineração de opinião para predição do comportamento do ativo da Petrobras (alta ou baixa), o qual apresentou bons resultados.

A extensão dos testes a outras duas empresas, obteve índice de acertos inferiores aos constatados durante a fase de treino e testes do referido modelo. No entanto, convém ressaltar que tais resultados constituem a fase inicial deste trabalho, cabendo algumas considerações que serão apresentadas adiante.

No capítulo que segue serão abordados alguns aspectos pertinentes a esta pesquisa, alguns deles foram observados e construídos ao longo deste trabalho, tais como considerações finais, contribuições, limitações, publicações alcançadas e, por fim, propostas para trabalhos futuros.

Capítulo 6

Conclusões

Neste capítulo serão apresentados aspectos conclusivos da pesquisa, como: considerações finais, contribuições, limitações, publicações e perspectivas para trabalhos futuros.

6.1 Considerações Finais

O mercado financeiro tem se revelado um ambiente bastante sensível a influências de fatores externos ao longo do tempo. A utilização do sentimento coletivo como variável que venha a melhorar a predição do comportamento das bolsas de valores se mostrou satisfatório, tendo como base as métricas do SVM.

Analisando-se o capítulo 5, pode-se perceber que a utilização das técnicas que envolvem o processo de descoberta do conhecimento, em especial a mineração de dados, teve grande importância para os resultados desta pesquisa.

O trabalho também evidenciou que os dados provenientes das postagens do *Twitter*, quando submetidos ao processo KDD, representam uma rica fonte de informação, em que a mineração de opinião é apenas uma das inúmeras utilizações que podem ser extraídas desses dados. No entanto, constatou-se que, para usufruir dos mesmos de forma satisfatória, torna-se necessário um grande esforço na etapa de pré-processamento do processo KDD, onde os dados são padronizados, pois a informalidade, característica desse tipo de comunicação, dificulta a utilização direta das técnicas posteriores à etapa de pré-processamento.

Outro fator que merece destaque é que eventualmente o humor coletivo não reflete o sentimento do mercado para aquele ativo específico. Tomando como exemplo o caso em estudo, a empresa brasileira Petrobras, que caracteriza por ser uma estatal de economia mista, como

já mencionado, vem passando por um momento de grandes incertezas, vendo-se envolvida em uma série de escândalos políticos e de desvio de dinheiro público.

Neste cenário desfavorável, boa parte dos comentários em que aparece o nome da empresa tem uma conotação essencialmente política, ou seja, há um desvio do foco nos comentários. Muitos usuários postam mensagens contendo insultos direcionados a políticos, por exemplo, o que certamente também aconteceria em outro cenário onde a política do país fosse mal e a empresa tivesse suas ações valorizadas, isto é, empresas com essa característica (estatal) podem em algum momento ter o humor coletivo distorcido pelo simples fato de terem sua imagem associada ao Estado.

O momento atual da empresa, com suas ações em crescente desvalorização, coincide com o momento político igualmente desfavorável, o que pode explicar os resultados apresentados no trabalho, que demonstram o humor predominantemente negativo nos dias em estudo, seguidos por uma desvalorização das ações.

Apesar dos estudos voltados para a mineração de opinião terem evoluído significativamente nos últimos anos, as ferramentas resultantes desse processo devem ser vistas como algo complementar no processo de tomada de decisão, diante da complexidade de extrair sentimentos com exatidão de fontes textuais em linguagem natural, como nesta pesquisa.

6.2 Contribuições

As principais contribuições deste trabalho, dentro de sua abordagem de predição utilizando mineração de opinião, foram as seguintes:

1. Avaliação da mineração de opinião, como uma abordagem que, quando combinada com técnicas de mineração de dados e aprendizagem de máquina, seja capaz de fornecer informações que tenham caráter preditivo satisfatório sobre os ativos de determinada empresa do mercado financeiro;

2. Ratificação do *Twitter* como sendo um "termômetro social", de maneira que possa retratar o sentimento social, partindo das mensagens postadas por seus usuários.

6.3 Limitações

Podem ser considerados como limitações da abordagem proposta os seguintes aspectos:

1. O modelo foi concebido utilizando como estudo de caso a empresa Petrobras, embora na fase de testes tenham sido adicionadas outras duas Empresas, não há conhecimento dos resultados quando utilizado um número significativo de empresas dos diferentes setores da economia no modelo;

2. O modelo fez uso de poucas amostras, no total de 55 instâncias, utilizadas para treinamento e testes do modelo. Fato que se repetiu também quando da utilização do protótipo para demais validações, principalmente ocasionado pela limitação na recuperação dos dados por parte da API do *Twitter*;

3. Grande dificuldade no processamento de linguagem natural, que é uma das bases para a mineração de opinião, utilizada neste trabalho.

6.4 Publicações

Esta pesquisa, durante seu curso, teve duas publicações aceitas sobre o tema proposto, sendo as duas de caráter internacional: em um jornal conferência (anexo I) e outra em uma conferência conferência (anexo J). Bibliografadas abaixo:

1. Lima, M. L. and Sofiane Labidi, Thiago P. do Nascimento and Nadson S. Timbó and Gilberto N. Neto and Marcus Vinicius Lima Batista (2016). *A Model Based on Sentiments Analysis for Stock Exchange Prediction – Case Study of PETR4, Petrobras, Brazil.* In *The Fourth International Conference on Artificial Intelligence, Soft Computing (AISC 2016)*, Zurich, Switzerland, January 02 03, 2016.

2. Lima, M. L. and Thiago P. Nascimento and Sofiane Labidi and Nadson S. Timbó and Marcos V. L. Batista and Gilberto N. Neto and Eraldo A. M. Costa and Sonia R. S. Sousa (2016). *Using Sentiment Analysis for Stock Exchange Prediction.* In *International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 7, No. 1, January 2016.*

6.5 Trabalhos Futuros

Durante as diversas fases que compõem a concepção de um trabalho científico, torna-se natural que surjam várias ideias que visem melhorar o seu propósito. Com este trabalho não foi diferente, citando-se a seguir propostas de estudos futuros:

- Tomando como base a mesma metodologia aplicada, expandi-la para empresas de portes diferentes como as *small caps*[1], que são empresas de baixo valor de mercado e estudar sua sensibilidade ao humor coletivo, ou seja, avaliar se o tamanho do capital da empresa é um fator determinante para este tipo de estudo;

- Segmentar a coleta das informações em espaços delimitados geograficamente, visando ter o sentimento coletivo por região e validar os resultados dessa nova abordagem, verificando se determinada área, tem maior peso preditivo que as demais dentro desse contexto;

- Criar um mecanismo, fazendo uso de ontologias, por exemplo, que impeça a interferência de postagens mal intencionadas, provenientes de robôs ou de forma manual que visem burlar o quantitativo e o teor natural das postagens, tentando formar uma opinião especulativa através de mensagens contendo boatos ou similares.

- Baseado na dificuldade do processamento de linguagem natural e visando obter melhores resultados na etapa de mineração de opinião, propor a criação um dicionário léxico próprio, específico para este domínio, contendo palavras e vocabulários técnicos utilizados no cotidiano do mercado financeiro e ponderando as palavras-chave, visando aumentar a confiabilidade de uma possível ferramenta.

[1] São empresas geralmente de médio e pequeno portes cuja negociação caracteriza-se pela descontinuidade e apresenta ações com pouca liquidez.

Referências Bibliográficas

Almeida, A. J. S. (2015). *Modelo de Predição para o Mercado Acionário Baseado na Lógica Fuzzy*. Dissertação de Mestrado, Universidade Federal do Maranhão.

Andreso, A. F. e Lima, I. S. (2007). *Mercado Financeiro: aspectos conceituais e históricos*. Atlas-Br, São Paulo, 3 edição.

Aranha, C. e Passos, E. (2007). Automatic nlp for competitive intelligence. *IGI Global Information Science Reference*.

Assaf Neto, A. (2006). *Finanças corporativas e valor*. Atlas, São Paulo, 2 edição.

Asur, S. e Huberman, A. (2010). Predicting the future with social media. *Web Intelligence and Intelligent Agent Technology (WI-IAT)*.

Bakshi, K. (2012). Considerations for big data: Architecture and approach. *Conference, 2012 IEEE, vol., no., pp.1,7, 3-10. doi: 10.1109/AERO.2012.6187357*.

Berkun, S. (2000). *The Art of UI Prototyping*. `http://www.scottberkun.com/essays/essay12.htm`, data de acesso: 20/05/2016.

Bernardo, I. S. P. B. (2014). *A Era de um Mercado Social: A Relação Entre o Twitter e o Mercado Acionista*. Dissertação de Mestrado – Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa.

Berry, M. J. e Linoff, G. (1997). Data mining techniques: for marketing, sales and customer support. USA: Wiley Computer Publishing.

BMFBovespa (2010). *Introdução ao mercado de capitais*. Manual referente aos mecanísmos e instumentos fornecidos pela Bolsa de Valores de São Paulo – Bovespa.

Bollen, J., Mao, H., e Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Booch, G., Rumbaugh, J., e Jacob, I. (1996). *UML Guia do Usuário*. Campus, 2000, 7a edição, Rio de Janeiro.

Bovespa (2000). *O Mercado de Capitais: Sua importância para o desenvolvimento e os entraves com que se defronta no Brasil*.

Cabena, P., P, H., e Stadler, R. (1997). *Discovering Data Mining From Concept To Implementation*. IBM. Data Mining, USA.

Capra, F. (2008). *O Tempo das Redes (Vivendo Redes)*. Editora Perspectiva.

Chorafas, D. N. (1992). *Treasury Operations & the Foreign Exchange Challenge*. John Wiley & Sons Ltda.

Cratochvil, A. (1999). *Data mining techniques in supporting decision making*. Master thesis – Universidade Federal do Rio Grande do Sul, Universiteit Leiden.

Cristianini, N. e Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Demirkan, H. e Delen, D. (2012). *Leveraging the capabilities of service-oriented decision support systems: Putting analytics and big data in cloud*. Decision Support Systems.

Dias, M. M. (2001). *Um Modelo De Formalização Do Processo De Desenvolvimento De Sistemas De Descoberta De Conhecimento Em Banco De Dados*. Tese de Doutorado - Universidade Federal de Santa Catarina, Santa Catarina-Brasil.

Dias da Silva, B. C. (2007). *Introdução ao Processamento das Línguas Naturais*. Núcleo Interinstitucional de Linguistica Computacional. NILC - ICMC-USP, São Paulo-Brasil.

Duarte, F. e Frei, K. (2008). *O Tempo das Redes (Redes Urbanas)*. Editora Perspectiva.

Emirbayer, M. e Goodwin, J. (1994). *Networ analysis, culture and the problem ofagency*. American Journal of Sociology, 99(6):1411–1454.

Fan, J. e Liu, H. (2013). *Statistical analysis of big data on pharmacogenomic*. Advanced Drug Delivery Reviews.

Fayyad, U. M., G., P.-S., e Smyth, P. (1996). *Advances in Knowledge Discovery and Data Mining*. AAAIPress, The Mit Press.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM, 56(4):82–89*.

Ferreira, G. C. (2011). Redes sociais de informação: uma história e um estudo de caso. *Perspect. ciênc. inf.*, 16(3):208–231.

Filho, S., da, L. A., Favero, E. L., e de, C. K. L. (2010). Mining association rules in data and text. *An Appplication In Public Sucurity. CONTECSI*.

Fortuna, E. (2008). *Mercado financeiro: Produtos e serviços*. Qualitymark - Br, Rio de Janeiro, 17 edição.

Gonzalez, M. e Lima, V. (2003). *Recuperação de Informação e Processamento da Linguagem Natural*. XXIII Congresso da Sociedade Brasileira de Computação. Anais da III Jornada de Mini-Cursos de Inteligência Artificial.Campinas. v. III. SP. Campinas, São Paulo-Brasil.

Graeml, A. R. (1998). *O valor da tecnologia da informação*. Anais do I Simpósio de Administração da Produção, Logística e Operações Industriais-EAESP-FGV, São Paulo.

Gupta, V. e Lehal, G. (2009). A survey of text mining techniques and applications. *Journal of Emerging Technologies in Web Intelligence, v. 1, n. 1 p. 60–76*.

Harrison, T. H. (1998). Internet data warehouse. São Paulo: Bekerley Brasil.

Helena, H. e Ângela C. (2003). Data mining concepts and applications. *Revista de Ciência & Tecnologia • V. 11, No 22 – pp. 19-34*.

Hu, M. e Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge*, páginas 22–25.

Java, A. (2007). *Why We Twitter: Understanding Microblogging Usage and Communities*. Joint 9th WEBKDD and 1st SNA-KDD Workshop, San Jose, California, USA.

Jurafsky, D. e Martin, J. H. (2000). *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, EUA : Prentice-Hall, 2000.

Lemos, R. e Santaella, L. (2010). *Redes sociais digitais: a cognição conectiva do Twitter*. Paulus.

Limeira, V. (2003). *Negócios em Bolsa de Valores: Estratégias para o Investimento*. Alaude.

Liu, B. (2010). Sentiment analysis and subjectivity. In *Handbook of Natural Language Processing, Second Edition. Taylor and Francis Group, Boca*.

Liu, B. (2012). Sentiment analysis and opinion mining.

Loh, S. (2001). *Abordagem Baseada em Conceitos para Descoberta de Conhecimento em Textos*. PhD thesis, Universidade Federal do Rio Grande do Sul, Instituto de Informática, Rio Grande do Sul.

Lopes, L. A. (2012). *Um sistema de inferência fuzzy como suporte a tomada de decisão para a compra e venda de ativos na bolsa de valores*. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) - Centro de Ensino Unificado de Teresina, Teresina.

Luquet, M. e Rocco, N. (2005). *Guia valor econômico de investimentos em ações*. Globo- Br, São Paulo.

Marques, F. C. R. (2010). *Maximização de Lucros em Investimentos: uma abordagem a partir do MACD com o emprego de algoritmos genéticos e lógica fuzzy*. Dissertação de Mestrado (Programa de Pós-Graduação em Modelagem Matemática e Computacional), Centro Federal de Educação Tecnológica de Minas Gerais, Belo Horizonte.

Mohammad, S. M., Kiritchenko, S., e Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA.

Nascimento, T. P. (2015). *Um Serviço Baseado em Algoritmos Genéticos para Predição da Bolsa de Valores*. Dissertação de Mestrado, Universidade Federal do Maranhão.

Neves, R. (2003). Pré-processamento no processo de descoberta de conhecimento em banco de dados. *UFRGS Programa de Pós-Graduação em Computador*.

Pal, S. K., Talwar, V., e Mitra, P. (2000). *Web Mining in Soft Computing Framework: Relevant, State of the Art and Future Directions*. Universiteit Leiden.

Palmisano, A. e Rosini, A. M. (2003). *Administração de Sistemas de Informação e a Gestão do Conhecimento*. Thomson. https://books.google.com.br/books?id=_t7D1uqWuUAC, data de acesso: 08/10/2015.

Pang, B. e Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval. doi:10.1561/1500000011.*

Pang, B., L. e Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. in proceedings of emnlp, pages 79–86.

Pender, T. (2004). *UML a Bíblia (p.42).* Campus Books, Rio de Janeiro-RJ.

Pinheiro, J. L. (2008). *Mercado de capitais: Fundamentos e técnicas*. Atlas - Br, São Paulo, 4 edição.

Popper, K. (1980). *A lógica da Investigação Científica. In: Os Pensadores.* Ed. Abril Cultural, São Paulo-SP.

Protalinski, E. (2013). *Twitter sees 218m monthly active users, 163.5m monthly mobile users, 100m daily users, and 500m tweets per day.* http://thenextweb.com/twitter/2013/10/03/twitter-says-it-sees-215-million-monthly-active-users-100, data de acesso: 20/01/2016.

Rezende, S., Pugliesi, J., Melinda, E., e Paula, M. (2003). *Mineração de Dados*. Editora Manole Ltda.

Rudd, J., Stern, K., e Isensee, S. (1996). *High-fidelity Prototyping Debate*. High-fidelity Prototyping Debate. Interactions, Vol.3, No 1, Janeiro de 1996.

Sandroni, P. (1994). *Dicionário de administração e finanças*. Ed Record - Br, São Paulo.

Santos, J. J. (2000). *Análise de custos: Remodelado com ênfase para custo marginal, relatórios e estudos de casos.* Atlas, São Paulo, 3 edição.

Souza, N. d. J. (2005). *Desenvolvimento Econômico*. Atlas - Br, São Paulo, 5 edição.

Statista (2016). Ranking das redes sociais por usuários registrados no mundo (em milhões). http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/, data de acesso: 01/02/2016.

Super, R. (2010). Brasil: país do twitter. *Revista Super Interessante, Ed.285, Dez/2010.*

Tan, P.-N., Steinbach, M., e Kumar, V. (2006). Introduction to data mining. *Addison Wesley.*

Tavares, M. e. M. F. (1987). *Análise Técnica - Avaliação de Investimentos.* IBMEC, Rio de Janeiro.

Thomsett, M. C. (1998). *Mastering Fundamental Analysis: how to spot trends and pick winning stocks using fundamental analysis.* Expert Systems with Applications, Chicago.

Tsytsarau, M. e Palmanas, T. (2012). Survey on mining subjective data on the web. In *Data Mining Knowledge Discovery, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 3, p. 478–514, may 2012. ISSN 1384-5810.*

Turney, P. D. (2002). semantic orientation applied tounsupervised classification of reviews.

Vapnik, V. (1995). The nature of statistical learning theory. *ACM SIGCOMM Computer Communication Review, v. 36, p. 7–15.*

Wiebe, J. M. (2002). semantic orientation applied tounsupervised classification of reviews.

Williams, N., Zander, S., e Armitage, G. (2006). A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification. *ACM SIGCOMM Computer Communication Review, v. 36, p. 7–15.*

Winger, B. e Frasca, R. (1995). Investments: Introduction to analysis and planning.

WU, X., ZHU, X., WU, G.-Q., e DING, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering, vol.26, no 1.*

Yeh, C.-Y., Huang, C.-W., e Lee, S.-J. (2011). *A multiple-kernel support vector regression approach for stock market price forecasting.* Expert Systems with Applications, 38 edição.

Zhang (2004). *Neural business in business forecasting.* Idea Group Publishing, São Paulo, 3 edição.

Zhang, X., Fuehres, H., e Gloor, P. (2010). Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Collaborative Innovation Networks Conference - COINs2010*, páginas 55–62.

Anexos

A Rotina Principal da Aplicação PHP para Coleta de Dados via API *Twitter*

```
<?php

session_start();
$_SESSION["maximoid"]=null;
include "trata_string.php";
require_once("twitteroauth/TwitterOAuth.php");
include "db_connect.php";
$contaSegundos = 0;

$twitteruser = "HDwAV2RVkhZKrs7hOk2VEroB5";
$notweets = 30;

$consumer = "HDwAV2RVkhZKrs7hOk2VEroB5";
$consumersecret ="2duTpVEg0Ck8XkZteV3rtSmwkwF5edPNSGJvechpgXoRIt2Vj8";
$accesstoken ="104881541-ZPRgeOZgclpyIvEQ0kzPZ0gUsvuEQBLVS2r9vPPa";
$accesstokensecret = "poSID3F9cJaAWwgdgFfMpIC9SPtYdnpwbUfbK9u5sqoxH";

$twitter = new
    TwitterOAuth($consumer,$consumersecret,$accesstoken,$accesstokensecret);

$del = $conn->exec("DELETE from tweets");
```

```
?>
  <html>
        <head>
              <meta charset="UTF-8" />
         </head>
        <body>
             <form action="" method="post" id="myform" name="myform">
             <label>Pesquisar: <input type ="text" name="keyword" value=""/></html>
             </form>
   <?php
   if (isset($_POST['keyword'])){
     for ($i = 1; $i <= 300000; $i++) {

          $contaSegundos++;
          echo $contaSegundos;
          if ($contaSegundos ==50) {
              sleep(60);
              $contaSegundos = 0;
          }

      if (isset($_SESSION["maximoid"])) {
        $strmax = '&max_id='. $_SESSION["maximoid"];
      } else {
        $strmax = '';
      }

     $strconsulta =
         'https://api.twitter.com/1.1/search/tweets.json?q=PETROBRAS&lang=pt&result_type=recent
         . $strmax;


         $tweets = $twitter->get($strconsulta);
         foreach($tweets->statuses as $tweet){
            $_SESSION["maximoid"]=$tweet->id_str;
           echo '<br>'.'<img src="'.$tweet->user->profile_image_url.'" />
               '.$_SESSION["maximoid"].'-'.$tweet->text;
```

```
            $novastring = remove_acentos($tweet->text);
                    $novaData = $tweet->created_at;
                    $LocalPost = $tweet->user->location;

          // new data
          $tweet = $novastring;
          $dia = $novaData;
          $local = $LocalPost;
          $id = $_SESSION["maximoid"];
          // query
                  $sql = "INSERT INTO tweets (id, dia, location, tweet) VALUES
                      ($id,:dia,:local,:tweet)";
                  $q = $conn->prepare($sql);
                  $q->execute(array(':dia'=>$dia,':local'=>$local,':tweet'=>$tweet));

           else {
             ;
          }

        }

     }
   }
?>
</body>
</html>
```

B Fragmento do Código Fonte onde Acontece a Autenticação com a API do *Twitter* Dentro da Aplicação PHP

```
<?php


session_start();
$_SESSION["maximoid"]=null;
include "trata_string.php";
require_once("twitteroauth/TwitterOAuth.php");
include "db_connect.php";
$contaSegundos = 0;


/*
USUARIO_TWITTER
*/
$twitteruser = "HDwAV2RVkhZKrs7hOk2VEroB5";
$notweets = 30;


/*
AUTENTICACAO
*/
$consumer = "HDwAV2RVkhZKrs7hOk2VEroB5";
$consumersecret ="2duTpVEg0Ck8XkZteV3rtSmwkwF5edPNSGJvechpgXoRIt2Vj8";
$accesstoken ="104881541-ZPRgeOZgclpyIvEQ0kzPZ0gUsvuEQBLVS2r9vPPa";
$accesstokensecret = "poSID3F9cJaAWwgdgFfMpIC9SPtYdnpwbUfbK9u5sqoxH";



$twitter = new
    TwitterOAuth($consumer,$consumersecret,$accesstoken,$accesstokensecret);
```

C Arquivo no Formato .CSV, Contendo os Atributos dos Indicadores Financeiros no período de 01/09/2015 a 20/11/2015.

```
/*
http://ichart.finance.yahoo.com/table.csv?s=PETR4.SA&a=08&b=01&c=2015&d=10&e=30&f=2015&g=d.,
*/
Date,Open,High,Low,Close,Volume,Adj Close
2015-11-30,7.54,7.84,7.42,7.67,72764300,7.67
2015-11-27,7.81,7.86,7.57,7.60,38334200,7.60
2015-11-26,7.95,8.04,7.82,7.92,21697100,7.92
2015-11-25,8.28,8.35,7.90,7.90,54732800,7.90
2015-11-24,8.05,8.50,7.95,8.50,63001700,8.50
2015-11-23,7.90,8.10,7.85,8.08,47415800,8.08
2015-11-20,7.84,7.84,7.84,7.84,000,7.84
2015-11-19,7.96,7.98,7.62,7.84,51878600,7.84
2015-11-18,7.81,8.03,7.79,7.82,38137700,7.82
2015-11-17,7.80,7.86,7.64,7.76,52072800,7.76
2015-11-16,7.31,7.70,7.31,7.70,56587800,7.70
2015-11-13,7.50,7.56,7.23,7.27,67000600,7.27
2015-11-12,7.72,7.74,7.53,7.58,41641400,7.58
2015-11-11,7.71,7.83,7.64,7.72,44596400,7.72
2015-11-10,7.60,7.69,7.51,7.61,34554300,7.61
2015-11-09,7.83,7.93,7.59,7.91,36340400,7.91
2015-11-06,8.04,8.08,7.73,7.82,39680600,7.82
2015-11-05,8.08,8.17,7.87,8.11,92690300,8.11
2015-11-04,8.58,8.65,8.02,8.08,55661500,8.08
2015-11-03,7.84,8.48,7.77,8.48,60306200,8.48
2015-11-02,7.71,7.71,7.71,7.71,000,7.71
2015-10-30,7.60,7.78,7.54,7.71,36951300,7.71
2015-10-29,7.57,7.93,7.57,7.61,41336300,7.61
2015-10-28,7.61,7.95,7.52,7.70,56685500,7.70
2015-10-27,7.86,7.86,7.55,7.60,44085700,7.60
2015-10-26,8.05,8.13,7.80,7.85,35069500,7.85
2015-10-23,8.26,8.28,7.94,7.99,45715300,7.99
2015-10-22,7.82,8.11,7.81,8.02,46174500,8.02
2015-10-21,7.95,8.06,7.71,7.75,45394400,7.75
```

2015-10-20,8.00,8.19,7.97,8.05,47938700,8.05

2015-10-19,7.90,7.96,7.76,7.95,40096200,7.95

2015-10-16,7.99,8.08,7.81,7.95,42963800,7.95

2015-10-15,8.00,8.07,7.67,8.00,61444100,8.00

2015-10-14,8.07,8.23,7.89,7.96,75339100,7.96

2015-10-13,8.36,8.47,8.07,8.13,56100600,8.13

2015-10-12,8.80,8.80,8.80,8.80,000,8.80

2015-10-09,9.02,9.07,8.58,8.80,71885200,8.80

2015-10-08,8.46,8.84,8.28,8.75,68670000,8.75

2015-10-07,8.51,8.90,8.27,8.47,116610600,8.47

2015-10-06,7.86,8.34,7.84,8.19,81259300,8.19

2015-10-05,7.90,8.14,7.75,7.82,68350300,7.82

2015-10-02,6.99,7.83,6.90,7.77,88891400,7.77

2015-10-01,7.31,7.46,6.99,7.02,69296500,7.02

2015-09-30,7.25,7.26,6.97,7.24,107779500,7.24

2015-09-29,6.54,6.79,6.50,6.59,52387600,6.59

2015-09-28,6.74,6.75,6.44,6.44,37466900,6.44

2015-09-25,7.12,7.22,6.77,6.82,44843100,6.82

2015-09-24,6.61,7.14,6.60,6.96,56577300,6.96

2015-09-23,6.96,7.16,6.70,6.82,64774800,6.82

2015-09-22,7.24,7.24,6.81,6.97,66720700,6.97

2015-09-21,7.66,7.68,7.30,7.30,41063600,7.30

2015-09-18,7.75,7.76,7.50,7.60,68097000,7.60

2015-09-17,8.05,8.11,7.86,7.86,72887600,7.86

2015-09-16,7.77,8.24,7.74,8.14,66613400,8.14

2015-09-15,7.65,7.87,7.54,7.65,35102900,7.65

2015-09-14,7.70,7.82,7.46,7.72,52851400,7.72

2015-09-11,7.94,7.99,7.59,7.66,53647300,7.66

2015-09-10,7.90,8.15,7.81,7.97,62858700,7.97

2015-09-09,8.81,8.90,8.36,8.39,38706300,8.39

2015-09-08,8.70,8.79,8.58,8.64,29383000,8.64

2015-09-07,8.51,8.51,8.51,8.51,000,8.51

2015-09-04,8.60,8.76,8.51,8.51,35463400,8.51

2015-09-03,8.93,9.21,8.54,8.76,56645400,8.76

2015-09-02,8.75,8.84,8.33,8.82,56050600,8.82

2015-09-01,8.75,9.09,8.54,8.59,52756400,8.59

D *Utilização Framework WEKA (fragmento do código)*

```
package br.ufma.lsi.weka;

import java.io.File;
import java.io.FileInputStream;
import java.io.InputStream;
import java.io.ObjectInputStream;
import java.net.URL;
import java.util.logging.Level;
import java.util.logging.Logger;
import weka.classifiers.Classifier;
import weka.classifiers.misc.InputMappedClassifier;
import weka.core.Instances;
import weka.core.SerializationHelper;
import weka.core.tokenizers.WordTokenizer;
import weka.experiment.InstanceQuery;
import weka.filters.Filter;
import weka.filters.unsupervised.attribute.Add;
import weka.filters.unsupervised.attribute.NominalToString;
import weka.filters.unsupervised.attribute.StringToWordVector;
import weka.filters.unsupervised.attribute.Reorder;


public class WekaClassifier {

    //Instances instances;
    public Instances stringToWordVector(Instances instances) {
        StringToWordVector stringToWordVector = null;
        // Set the tokenizer
        WordTokenizer wordTokenizer = new WordTokenizer();
        wordTokenizer.setDelimiters("\r\n\t.,;:\"\"()?!");

        try {
            stringToWordVector = new StringToWordVector();
```

```
            stringToWordVector.setInputFormat(instances);
            stringToWordVector.setOptions(weka.core.Utils.splitOptions(" -R
                first-last -W 1000 -prune-rate -1.0 -T -I -N 0 -stemmer
                weka.core.stemmers.NullStemmer -M 1 -tokenizer
                \"weka.core.tokenizers.WordTokenizer -delimiters \\\"
                \\\\r\\\\n\\\\t.,;:\\\\\\\'\\\\\\\"()?!\\\"\""));
            instances = Filter.useFilter(instances, stringToWordVector);
            //System.out.println(instances.toString());
        } catch (Exception ex) {
            Logger.getLogger(WekaClassifier.class.getName()).log(Level.SEVERE, null,
                ex);
        }
        return instances;
    }

    public Instances reorder(Instances instances) {
        Reorder reorder = new Reorder();
        try {
            reorder.setInputFormat(instances);
            reorder.setOptions(weka.core.Utils.splitOptions(" -R first-last"));
            instances = Filter.useFilter(instances, reorder);
            reorder.setOptions(weka.core.Utils.splitOptions(" -R last-first"));
            instances = Filter.useFilter(instances, reorder);
            instances.setRelationName("database_tweets");
            //System.out.println(instances.toString());
        } catch (Exception ex) {
            Logger.getLogger(WekaClassifier.class.getName()).log(Level.SEVERE, null,
                ex);
        }
        return instances;
    }

    public Instances loadDataBase() {
        Instances instances = null;
        InstanceQuery query;
        try {
```

```
        URL url;
        url = getClass().getResource("DatabaseUtils.props");
        query = new InstanceQuery();
        query.initialize(new File(url.getPath()));
        query.setDatabaseURL("jdbc:mysql://localhost:8889/TwitterMessage");
        query.setUsername("root");
        query.setPassword("root");
        String sql = "SELECT stocks.openstock abertura, stocks.low minomo,
            stocks.high maximo, stocks.closestock fechamento, "
                + "if(IFNULL(tab1.quant,0) > IFNULL(tab2.quant,0), 'NEGATIVO',
                    'POSITIVO') sentimento "
                + "FROM (select COUNT(corpus.tweetPolarity) quant,
                    corpus.tweetPolarity, DATE_FORMAT(tweets.createdAt,
                    '%Y-%m-%d') createdat from corpus, tweets "
                + "where tweets.tweetID=corpus.tweetID "
                + "and corpus.tweetPolarity=0 "
                + "group by corpus.tweetPolarity, DATE_FORMAT(tweets.createdAt,
                    '%Y-%m-%d') "
                + "order by DATE_FORMAT(tweets.createdAt, '%Y-%m-%d') "
                + ") tab1 "
                + "LEFT JOIN (select COUNT(corpus.tweetPolarity) quant,
                    corpus.tweetPolarity, DATE_FORMAT(tweets.createdAt,
                    '%Y-%m-%d') createdat from corpus, tweets "
                + "where tweets.tweetID=corpus.tweetID "
                + "and corpus.tweetPolarity=4 "
                + "group by corpus.tweetPolarity, DATE_FORMAT(tweets.createdAt,
                    '%Y-%m-%d') "
                + "order by DATE_FORMAT(tweets.createdAt, '%Y-%m-%d') "
                + ") tab2 ON tab1.createdat=tab2.createdat, stocks "
                + "WHERE DATE_FORMAT(stocks.createdat,
                    '%Y-%m-%d')=IFNULL(tab1.createdat, tab2.createdat)";
        query.setQuery(sql);
        query.close();
        instances = query.retrieveInstances();
    } catch (Exception ex) {
```

```
            Logger.getLogger(WekaClassifier.class.getName()).log(Level.SEVERE, null,
                ex);
        }
        return instances;
    }

    public Instances nominalToString(Instances instances) {
        NominalToString nominalToString = new NominalToString();
        try {
            nominalToString.setInputFormat(instances);
            instances = Filter.useFilter(instances, nominalToString);
        } catch (Exception ex) {
            Logger.getLogger(WekaClassifier.class.getName()).log(Level.SEVERE, null,
                ex);
        }
        return instances;
    }

    public Instances addClass(Instances instances) {
        Add add = new Add();
        try {
            add.setInputFormat(instances);
            add.setOptions(weka.core.Utils.splitOptions(" -T NOM -N class -L
                \"ALTA,BAIXA\" -C last"));
            instances = Filter.useFilter(instances, add);
            instances.setClassIndex(1);
        } catch (Exception ex) {
            Logger.getLogger(WekaClassifier.class.getName()).log(Level.SEVERE, null,
                ex);
        }
        return instances;
    }

    public Classifier getInputMappedClassifier(String name) {
        InputMappedClassifier inputMappedClassifier = new InputMappedClassifier();
        String pathModel = getClass().getResource(name + ".model").getFile();
```

```
        try {
            inputMappedClassifier.setOptions(weka.core.Utils.splitOptions(" -I -trim
                -L " + pathModel + " -W weka.classifiers.rules.car.WeightedClassifier
                -- -A \"weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M
                0.1 -S -1.0 -c -1\" -L -1 -W Equal"));
        } catch (Exception ex) {
            Logger.getLogger(WekaClassifier.class.getName()).log(Level.SEVERE, null,
                ex);
        }
        return inputMappedClassifier;
    }

    public Classifier getClassifier(String name) {
        Classifier classifier = null;
        InputStream inputStream;
        inputStream = getClass().getResourceAsStream(name + ".model");
        try {
            classifier = (Classifier) SerializationHelper.read(inputStream);
        } catch (Exception ex) {
            Logger.getLogger(WekaClassifier.class.getName()).log(Level.SEVERE, null,
                ex);
        }
        return classifier;
    }

    public Classifier loadModel(File path, String name) throws Exception {
        Classifier classifier;
        FileInputStream fis = new FileInputStream(path + name + ".model");
        ObjectInputStream ois = new ObjectInputStream(fis);
        classifier = (Classifier) ois.readObject();
        ois.close();
        return classifier;
    }
}
```

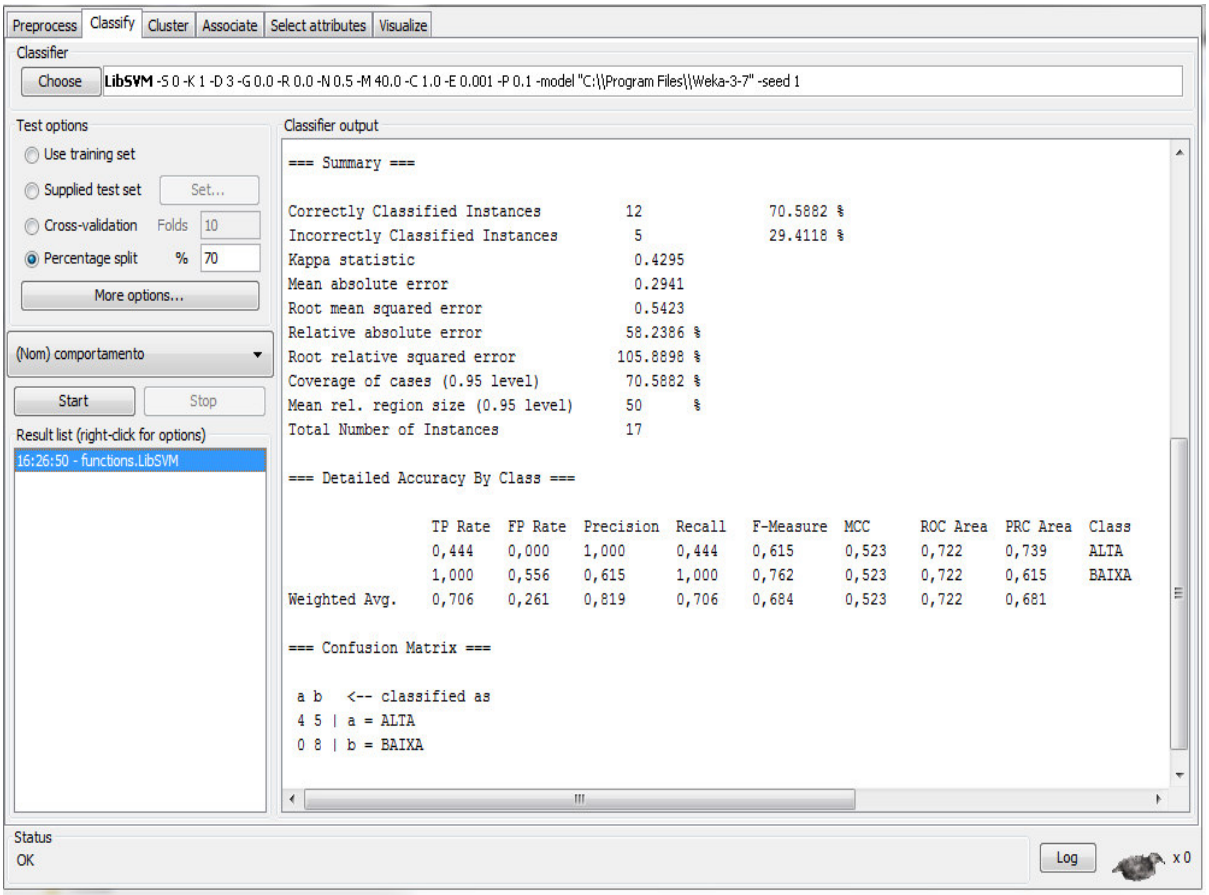
E *Acesso a API do Sentimento140 (fragmento do código)*

```
<?php
$data_string= $data_string . "]}";
//print_r($data_string);
$ch = curl_init() ;
curl_setopt($ch, CURLOPT_URL, "http://www.sentiment140.com/api/
bulkClassifyJson?appid=milsonlima@hotmail.com/");
curl_setopt($ch, CURLOPT_CUSTOMREQUEST, "POST");
curl_setopt($ch, CURLOPT_POSTFIELDS, $data_string);
curl_setopt($ch, CURLOPT_RETURNTRANSFER, true);
curl_setopt($ch, CURLOPT_HTTPHEADER, array(
    'Content-Type: application/json; charset=utf-8',
    'Content-Length: ' . strlen($data_string))
);
//send tweets to Sentiment140 using curl and store result in
variable $result
$result = curl_exec($ch) . "";
echo "<br>";
//write results to array - Had to use utf8-encode as getting
spurious results from result.
$json=json_decode(utf8_encode($result));
$count=0;
$i=0;

//Loop para inserção
foreach ($json->data as $polItems) {
    if ($i>0) {
        $sql=$sql.",";
    } else {
        $sql="INSERT INTO tweets (tweetSubject, tweetID, tweetText,
createdAt) values ";
    }
        $sql=$sql . "(\"" . $inputSearch . "\",\"" . $polItems->id .
"\",\"" . $list[$i] .  "\",'" .  date("Y-m-d H:i:s",
strtotime($polItems->date)) . "')";
    $i++;
}
//Corpus com polaridade definida
foreach ($json->data as $polItems) {
    if ($count>0) {
        $sql1=$sql1.",";
    } else {
        $sql1="INSERT INTO corpus (tweetID, tweetText,
tweetPolarity) values ";
    }
        $sql1=$sql1 . "(\"" . $polItems->id . "\",\"" . $polItems-
>text . "\"," .  $polItems->polarity . ")";
    $count++;
}
```

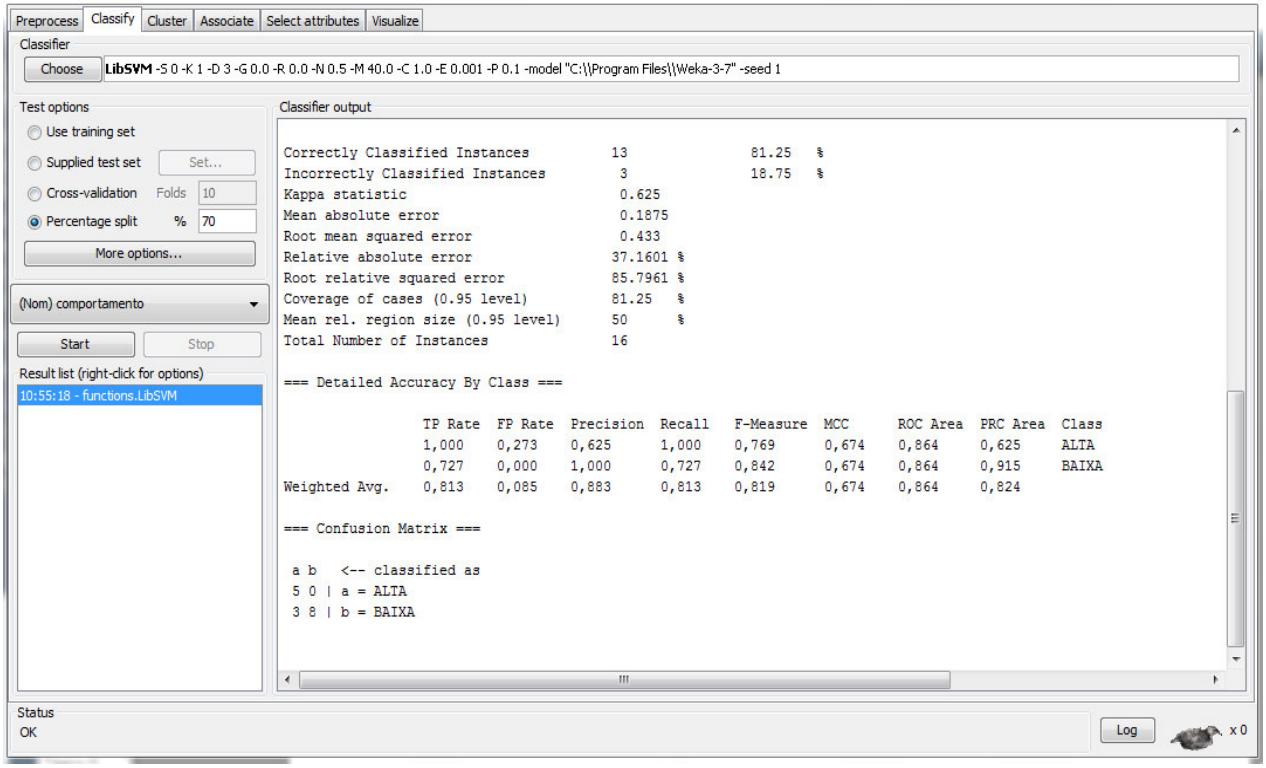
F *Estrutura do Banco de dados para o Protótipo*

```
/*
 Source Server         : Local
 Source Server Type    : MySQL
 Source Server Version : 50542
 Source Host           : localhost
 Source Database       : TwitterMessage
*/

--  Table structure for `corpus`
DROP TABLE IF EXISTS `corpus`;
CREATE TABLE `corpus` (
  `tweetID` bigint(20) NOT NULL DEFAULT '0',
  `tweetText` varchar(140) COLLATE utf8_unicode_ci DEFAULT NULL,
  `tweetPolarity` tinyint(4) DEFAULT NULL,
  PRIMARY KEY (`tweetID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;

--  Table structure for `stocklist`
DROP TABLE IF EXISTS `stocklist`;
CREATE TABLE `stocklist` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `stockcode` varchar(30) NOT NULL,
  `stocktext` varchar(30) NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB AUTO_INCREMENT=5 DEFAULT CHARSET=utf8;

--  Table structure for `stocks`
DROP TABLE IF EXISTS `stocks`;
CREATE TABLE `stocks` (
  `id` bigint(20) unsigned NOT NULL AUTO_INCREMENT,
  `openstock` decimal(10,2) NOT NULL,
  `low` decimal(10,2) NOT NULL,
  `high` decimal(10,2) NOT NULL,
  `closestock` decimal(10,2) NOT NULL,
  `createdat` datetime NOT NULL,
  `kind` varchar(5) DEFAULT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

--  Table structure for `stopWords`
DROP TABLE IF EXISTS `stopWords`;
CREATE TABLE `stopWords` (
  `stopWord` varchar(255) NOT NULL,
  `id` int(11) NOT NULL AUTO_INCREMENT,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB AUTO_INCREMENT=526 DEFAULT CHARSET=latin1;

--  Table structure for `tweets`
DROP TABLE IF EXISTS `tweets`;
CREATE TABLE `tweets` (
  `tweetSubject` varchar(255) COLLATE utf8_unicode_ci NOT NULL,
  `tweetID` bigint(20) NOT NULL DEFAULT '0',
  `tweetText` varchar(255) COLLATE utf8_unicode_ci NOT NULL,
  `createdAt` datetime NOT NULL,
  PRIMARY KEY (`tweetID`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8 COLLATE=utf8_unicode_ci;

--  Table structure for `users`
DROP TABLE IF EXISTS `users`;
CREATE TABLE `users` (
  `id` int(11) NOT NULL AUTO_INCREMENT,
  `username` varchar(50) CHARACTER SET utf8 COLLATE utf8_unicode_ci NOT NULL,
  `password` varchar(16) CHARACTER SET utf8 COLLATE utf8_unicode_ci NOT NULL,
  PRIMARY KEY (`id`)
) ENGINE=InnoDB AUTO_INCREMENT=4 DEFAULT CHARSET=utf8;
```

G Tela do Weka Mostrando os Resultados para o Modelo com Utilização Somente dos Indicadores Financeiros (Etapa-1)



H Tela do Weka Mostrando os Resultados para o Modelo com a Inserção do Atributo “Sentimento” (Etapa - 2)



I *Using Sentiment Analysis for Stock Exchange Prediction.* In *International Journal of Artificial Intelligence & Applications (IJAIA), Vol. 7, No. 1, January 2016*

International Journal of Artificial Intelligence & Applications (IJAIA)

http://airccse.org/journal/ijaia/ijaia.html

21/01/2016

To,

Milson L.Lima,
Post-Graduation Program in Electrical Engineering,
Federal University of Maranhão,
MA, Brazil.

Dear Sir,

On behalf of the Editorial Board of **International Journal of Artificial Intelligence & Applications (IJAIA)**, I am glad to inform you that based on the recommendations made by the reviewers on the paper submitted by you entitled: "**Using Sentiment Analysis for Stock Exchange Prediction (Milson L.Lima, Thiago P.Nascimento, Sofiane Labidi, Nadson S.Timbó, Marcos V.L.Batista, Gilberto N.Neto, Eraldo A.M.Costa and Sonia R.S.Sousa)**", has been accepted for **January 2016,, Volume 7, Number 1** publication in IJAIA.

Secretary, AIRCC

AIRCC Publishing Corporation

www.airccse.org

J *A Model Based on Sentiments Analysis for Stock Exchange Prediction – Case Study of PETR4, Petrobras, Brazil.* In *The Fourth International Conference on Artificial Intelligence, Soft Computing (AISC 2016)*, Zurich, Switzerland, January 02 03, 2016.

Fourth International Conference on Artificial Intelligence, Soft Computing

(AISC-2016)

January 02~03, 2016, Zurich, Switzerland

http://ccsit2015.org/2016/aisc/index.html

19/12/2015

Dear Milson L.Lima,

First of all, thank you very much for submitting your paper to **AISC 2016** to be held in **Zurich, Switzerland, January 02~03, 2016**. Based upon the reviewer's reports, we are pleased to inform you that your paper titled "**A Model Based on Sentiments Analysis for Stock Exchange Prediction - Case Study of Petr4, Petrobras, Brazil (Milson L. Lima, Sofiane Labidi, Thiago P. do Nascimento, Nadson S. Timbó, Gilberto N. Neto, Marcus Vinicius Lima Batista)**" has been ACCEPTED by the conference and will be included in the proceedings published by Computer Science Conference Proceedings in Computer Science & Information Technology (CS & IT) series. Congratulations on your excellent work!

In order to achieve the highest quality proceedings, we urge you to carefully consider the reviewer's comments, if any, when preparing the final version of your paper.

1. Please read the following Information carefully to prepare a final manuscript of your paper

http://airccse.org/journal/aircc_template.doc Maximum number of pages without extra payment is 20 (CCSP format). For each extra page you have pay 50 USD additionally.

2. Submit your final camera ready version of paper (.doc version+ .pdf version) and filled CR form **aisc_conf@ccsit2015.org (or) aisc_conf@yahoo.com**

3. Final Manuscript Submission Details:

a) When submitting your final manuscript, please ensure that you send us all source files such as .doc and pdf.

Copy right form: Volume editors: **David Wyld et al.**,

b) Please make sure you enter volume editor's name (David Wyld et al.,) in the copy right form.

Registration Details:

For International Authors

Regular Registration: **350 EURO**

For International Authors

Payment Methods: wire transfer/Bank transfer/Net transfer

Please note that at least one of the authors of accepted papers is required to register and present the paper at the conference;

Thank you again for helping to ensure the success of AISC **2016**. We are looking forward to meeting you at the **AISC 2016** in **Zurich, Switzerland, January 02 ~ 03, 2016**.

Best regards,
Secretary, AISC 2016
{For} Program Chairs/Organizers
http://ccsit2015.org/2016/aisc/index.html

Secretary, AISC 2016